이학박사 학위논문

Reliable Application of Neural-Network to Astrophysical Problems: Generation of Gravitational Waveforms and Photometric Redshifts of Galaxies

천체물리학 문제들에 대한 인공신경망의 고신뢰적 적용: 중력파 파형 생성과 은하 측광 적색편이

2021년 8월

서울대학교 대학원

물리·천문학부 천문학전공

이 준 구

Reliable Application of Neural-Network to Astrophysical Problems: Generation of Gravitational Waveforms and Photometric Redshifts of Galaxies

천체물리학 문제들에 대한 인공신경망의 고신뢰적 적용: 중력파 파형 생성과 은하 측광 적색편이

지도교수 이형목

이 논문을 이학박사 학위논문으로 제출함

2021년 7월

서울대학교 대학원

물리·천문학부 천문학전공

이 준 구

이 준 구의 이학박사 학위논문을 인준함

2021년 8월

위 원 장 _____

부 위 원 장 _____

위 원 _____

위 원 _____

위 원 _____

Reliable Application of Neural-Network to Astrophysical Problems: Generation of Gravitational Waveforms and Photometric Redshifts of Galaxies

by

Joongoo Lee
(jglee@astro.snu.ac.kr)

A dissertation submitted in partial fulfillment of the requirements for
the degree of

Doctor of Philosophy

in
Astronomy

in
Astronomy Program
Department of Physics and Astronomy
Seoul National University

Committee:

Professor	Myungshin Im
Professor	Hyung Mok Lee
Professor	Myung Gyoon Lee
Professor	Min-Su Shin
Professor	Sang Hoon Oh

To my family and friends.

ABSTRACT

Neural network (NN) is one of the representative machine learning algorithms showing remarkable performances in a wide variety of tasks and fields. However, the valid application range and credibility of the NNs are still questionable. In this thesis, we 1) present that NN may achieve greater or comparable performances compared to conventional methods for gravitational waveform generation and photometric redshift estimation, 2) study how the performance of NNs variates with respect to data property, 3) show the performance degradation of NNs for test data drawn from different distributions from those of training data. Numerous accurate waveform templates are necessary for the precise detection of gravitational waves with small strain amplitude compared to observed noise, and the templates can be accurately computed by numerical relativity. However, numerical relativity is not capable of computing waveforms in a short time because the method requires a huge amount of computation resources. We introduce a novel NNs based on recurrent NNs, which can compute $\sim$ 1500 waveforms in $O(1)$ second with accuracy over 99%. To validate its practicality, we use parameter estimation using an artificial signal injected into real noise data from Laser Interferometer Gravitational-Wave Observatory. In addition, we prove that the application of the NNs is not limited to one astronomical task by applying NNs to photometric redshift estimation of galaxies. The model produces highly accurate photometric redshifts only using color-related features as inputs. Simultaneously, we also show that the NNs accuracy rapidly diminishes for the data residing in the low-density region in the input space and unseen data during training, i.e., quasars and stars. For practical application of the NNs to astrophysical problems, we define these data as out-of-distribution (OOD) and design a method using unsupervised training capable of detecting OOD with accuracy over 98% while maintaining the performance of the photometric redshift estimation.

Keywords: machine – learning: neural – network: gravitational – waves: galaxies: quasars: stars: Out-of-Distribution

Student Number: 2014-22385

Contents

[1] Lee et al. Phys. Rev. D 103, 123023 (2021)

[2]Submitted to the Astronomical Journal

[3]To be submitted to the Astronomical Journal

List of Figures

List of Tables

Chapter 1

Introduction

We are now in the midst of an artificial intelligence renaissance driven by big data. Artificial intelligence refers to the incarnation of intelligence in machines that are programmed to think like humans and imitate their actions. One of the pioneering studies on the modern concept of artificial intelligence is the logical framework of, a British mathematician, Alan Turing's paper published in 1950 (Turing 1950). Through this one of the very first papers in the field of artificial intelligence, he discussed how to build intelligent machines and suggested the so-called Turing test to test their intelligence. Since the first introduction of the Turing test, it has become a nucleus concept in the philosophy of artificial intelligence.

Machine learning, seen as a part of artificial intelligence, is the study of computer algorithms that automatically improve through experience and accumulated data. The term machine learning was coined in 1959 by Arthur Samuel, an American pioneer of artificial intelligence (Samuel 1959) who invented one of the world's first successful self-learning programs, the Samuel Checkers-playing Program. Since the invention of the first self-teaching system, manifold breakthroughs in machine learning have arisen with the invention of plenty of machine learning models, such as K-nearest neighbors (Altman 1992), support vector machine (Cortes & Vapnik 1995), decision tree (Quinlan 1986), random forest (Breiman 2001), etc.

The multiple breakthroughs in machine learning have enabled extensive researches using big data in modern computer science. Of all the fruits rooted in the breakthroughs, a neural

network inspired by the functioning of the human brain (Hopfield 1982) is the most popularly used machine learning architecture. There are many types of neural networks: fully connected networks, convolutional neural networks (Fukushima 1980), recurrent neural networks (Rumelhart et al. 1986), etc. These variations of neural networks are used for different purposes and data.

The fully connected networks are widely used for feature-based data, that the order of each feature has no effect on the significance of the data. Generally, the fully connected networks consist of one input layer, multiple hidden layers, and one output layer. These layers are again composed of artificial neurons or perceptrons (Rosenblatt 1958), which are the basic units forming the model and mimicking biological neurons. The artificial neurons are interconnected to the ones in adjacent layers with randomly initialized connection weights. Vectors given to the neurons in the input layer are transformed by the neurons in the hidden layers using connection weights with non-linear activation functions and are transmitted to the output layer.

The convolutional neural networks are optimized for handling image type data that the positions of each pixel are dependent. The architecture of the networks, inspired by the visual cortex, is similar to the connectivity pattern of neurons in the human brain. The networks comprise one or more convolutional layers and fully connected layers on top and/or bottom. The convolution operated by convolutional layers with a collection of tied weights, so-called kernel, results in an activation map (or feature map) showing the strength of the detected features and their positions in the provided input image. To reduce the spatial size of the convolved feature map, convolutional networks use pooling layers, extracting some dominant features from the map.

The recurrent neural networks are designed for the application of time-series data. The recurrent networks consist of recursively called neural networks, allowed to use an element of the sequential data with the previous activations of the hidden layers as its inputs. The characteristics of the recurrent network allow the model to remember its past activations and make a decision influenced by what it has learnt from the past. To enhance the networks' memory for important past for its decision, improved networks such as long short-term memory networks (Hochreiter & Schmidhuber 1997) and gated recurrent units (Cho et al. 2014) are

introduced.

The neural networks and their variations are renowned for its capability of handling a massive amount of data and shows remarkable performances in a wide variety of fields and tasks: for example, image classification (Russakovsky et al. 2015; Krizhevsky et al. 2017, 2012), autonomous vehicles (Levinson et al. 2011; Dosovitskiy et al. 2017b), protein structure prediction (Senior et al. 2020; Torrisi et al. 2020), and natural language processing (Sutskever et al. 2014; Kalchbrenner & Blunsom 2013). At present, neural network as a research tool penetrates the field of science, crossing the barrier of research areas. Astronomy is also not an exception.

Plural ongoing and past studies in Astronomy have shown that the neural network may achieve comparable or even superior performances to the conventional methods for various astronomical tasks. The example research areas include galaxy evolution (Ghosh et al. 2020; Reiman 2020), extragalactic and cosmological studies (Perraudin et al. 2019; Piscopo et al. 2019), gravitational waves (J. Lee et al. 2021), active galactic nuclei (Ellison et al. 2016; Chen et al. 2021), dark matter and dark energy surveys (Berger & Stein 2019; Escamilla-Rivera et al. 2020), etc.

However, the valid application range and credibility of the neural network are still questionable. For example, the stable performance of the networks is guaranteed only when the tested samples are drawn from the same distribution as training data (Hendrycks & Gimpel 2016). In practice, the assumption that all of the samples in real-world data are well-controlled and come from the same distribution as the training data is implausible. In this case, the performance of the networks deteriorates, and the model becomes untrustworthy (Liang et al. 2017; Ren et al. 2019; Yu & Aizawa 2019). Hence, the exhaustive validity scrutiny of the trained networks on the samples drawn from out-of-distribution (OOD) with respect to training data is necessary for more reliable usage of the neural network.

Throughout this thesis, we introduce multiple approaches using neural networks for astronomical application and examine the validity of the trained networks using diverse verification. In Chapter 2, we present a novel architecture based on a recurrent neural network for generating gravitational waveforms from binary black hole coalescences. We validate the model for practical usage with parameter estimation using injected signal into a real advanced Laser

Interferometer Gravitational-wave Observatory (aLIGO) data. Chapter 3 is devoted to neural network inference for the photometric redshifts of galaxies observed by Panoramic Survey Telescope and Rapid Response System (Pan-STARRS) surveys. In this chapter, we offer an in-depth exploration of the valid boundaries of model application. We proceed with follow-up research of Chapter 3 in Chapter 4, which investigates the influence of OOD objects on model performance. We also introduce a way to flag OOD examples as they are given to the model.

Chapter 2

Deep Learning Model on Gravitational Waveforms in Merging and Ringdown Phases of Binary Black Hole Coalescences[1]

Abstract

The waveform templates of the matched filtering-based gravitational-wave search ought to cover wide range of parameters for the prosperous detection. Numerical relativity (NR) has been widely accepted as the most accurate method for modeling the waveforms. Still, it is well-known that NR typically requires a tremendous amount of computational costs. In this chapter, we demonstrate a proof-of-concept of a novel deterministic deep learning (DL) architecture that can generate gravitational waveforms from the merger and ringdown phases of the non-spinning binary black hole coalescence. Our model takes $O(1)$ seconds for generating approximately 1500 waveforms with a 99.9% match on average to one of the state-of-the-art waveform approximants, the effective-one-body. We also perform matched filtering with

[1] Lee et al. Phys. Rev. D 103, 123023 (2021)

the DL-waveforms and find that the waveforms can recover the event time of the injected gravitational-wave signals.

2.1 Introduction

Since the first detection of gravitational waves (GW)(Abbott et al. 2016), numerous GW events have been captured by ground-based GW detectors, the Advanced Laser Interferometer Gravitational-wave Observatory (aLIGO) (Aasi et al. 2015) and Virgo (Acernese et al. 2015). The sources of all events turned out to be compact binary coalescences (CBCs), the collision of two dense objects such as black holes (BH) or neutron stars (NS) — mostly from binary black holes (BBH), 47 out of 50, and partially from binaries containing at least one neutron star (Abbott et al. 2020).

For the type of GW progenitors, template-based GW search is one of the most efficient approaches because the gravitational waveforms from binary mergers can be modeled precisely by multiple methods, e.g., post-Newtonian (PN) for the inspiral phase, numerical relativity for the merger phase, and perturbation theory for the ringdown phase. The template-based search utilizes the matched filtering method (Turin 1960), which essentially computes the cross-correlation between template waveforms and real GW signal buried in noisy data.

The successful implementation of the matched-filtering-based search relies on the pre-computed waveform templates. Numerical relativity (NR) has been considered as the most accurate method for computing gravitational waveforms. However, obtaining a large number of templates that cover parameter space densely enough for the precise matched filtering search and parameter estimation with NR is not feasible because of too heavy computational requirements. For example, NR simulation of the first GW event GW150914 (Abbott et al. 2016) takes 1-2 weeks using tens to hundreds of CPU cores (Lovelace et al. 2016). In contrast, it takes less than $O(1)$ seconds to generate inspiral waveforms using post-Newtonian approximations.

Several waveform models approximating NR waveforms have been proposed to reduce the computational cost with reasonable accuracy NR (Blanchet et al. 1995; Droz et al. 1999; Buonanno & Damour 2000; Blanchet et al. 2004; Pürrer 2014; Taracchini et al. 2014; Pürrer 2016; Bohé et al. 2016). Nonetheless, the physical parameter spaces where each approximant exactly

covers are different from each other (Kumar et al. 2016; Taracchini et al. 2014). Therefore, reserving plural waveform models, complementing each other for various configurations, and saving computing time are crucial for a more elaborate template-based search. It justifies the further study of new waveform approximants.

We present a proof-of-concept demonstration of a deep learning (DL) model for generating gravitational waveforms from the CBC events covering the late phase of inspiral to final ringdown phases. For this purpose, we only consider non-spinning BBH systems for simplicity. Chua et al. (Chua et al. 2019) utilize deep artificial neural networks to map the physical parameters to coefficients of reduced-order bases waveforms. Williams et al. (Williams et al. 2019) use Gaussian process regression to approximate the inspiral-merger-ringdown waveforms from the BBH. However, the capability of a fully DL-based deterministic approach has not been explored so far for the generation of the merger-ringdown waveform of CBC[2]. Hence, we examine the viability of the deterministic DL model as a merger-ringdown gravitational waveform model throughout this chapter.

While DL models show remarkable performances in a wide variety of fields such as natural language processing (NLP) (Brown et al. 2020; Vaswani et al. 2017), autonomous driving (Dosovitskiy et al. 2017a), and image classification (Krizhevsky et al. 2017), most of the models are only capable of handling fixed-size data once they are trained. However, the model we shall adopt for this study should be able to cope with differently-sized data because the length of the waveforms observable by GW detectors depends on the two factors: (1) lower-frequency limit of the detector's sensitivity (around 10 Hz for ground-based detectors) and (2) the masses of the compact binary system (Aasi et al. 2015; Abbott et al. 2020).

The recurrent neural network (RNN) encoder-decoder-based sequence-to-sequence (seq2seq) model (Cho et al. 2014; Sutskever et al. 2014) designed for NLP is one of the DL models that can handle variable input/output sizes. This model also has shown outstanding performances in many NLP studies (Gehring et al. 2017; Venugopalan et al. 2015; Luong et al. 2015; Nallapati et al. 2016). The property of gravitational waveforms is similar to that of language type data

[2]It is known that deterministic models generally show higher accuracy and performance than stochastic methods as the training data is sufficient.

Table 2.1 Parameters of the waveform in each dataset. Dataset-1 and-2 have different mass ranges, mass ratios, and numbers of samples, as shown in the table. All the other parameters of both datasets are set to be the same. Note that the waveforms in the datasets are generated in the time domain with `PyCBC` and `SEOBNRv4`.

Variable	Dataset-1	Dataset-2
Mass [min, max]	$[10M_\odot, 40M_\odot]$	$[40M_\odot, 100M_\odot]$
Mass ratio [min, max]	[1, 4]	[1, 2.5]
Number of waveforms (training, validation, test)	(12469, 1533, 1512)	(12447, 1530, 1523)
sampling rate	4096Hz	4096Hz
Distance	100Mpc	100Mpc
Spin	0	0
Inclination angle	30°	30°

containing time-ordered words in sentences with different lengths. In that sense, we consider seq2seq as the experimental method to generate waveforms and slightly modify the structure of the model for our purpose.

This chapter is organized as follows. Sec 2.2 provides detailed explanations on the data preparation. In Sec 2.3, the original seq2seq model, our modified version, and an evaluation metric for the model performance are elaborated. Sec 2.4 presents the results of the DL-waveform analysis with GW data and additional dataset-size-associated experiments. Finally, we discuss our results and future research directions in Sec 2.5.

2.2 Data

Since RNN is well-suited to time-series data, we compute non-spinning BBH waveforms in time-domain for training dataset using `PyCBC` (Nitz et al. 2018), a software package for GW data analysis. For this, we use a variant of effective-one-body (EOB) approximants, `SEOBNRv4` (Bohé et al. 2017), one of the most accurate versions of the approximants used in the GW searches.

For the training of the DL model, adopting waveforms obtained by NR is more beneficial

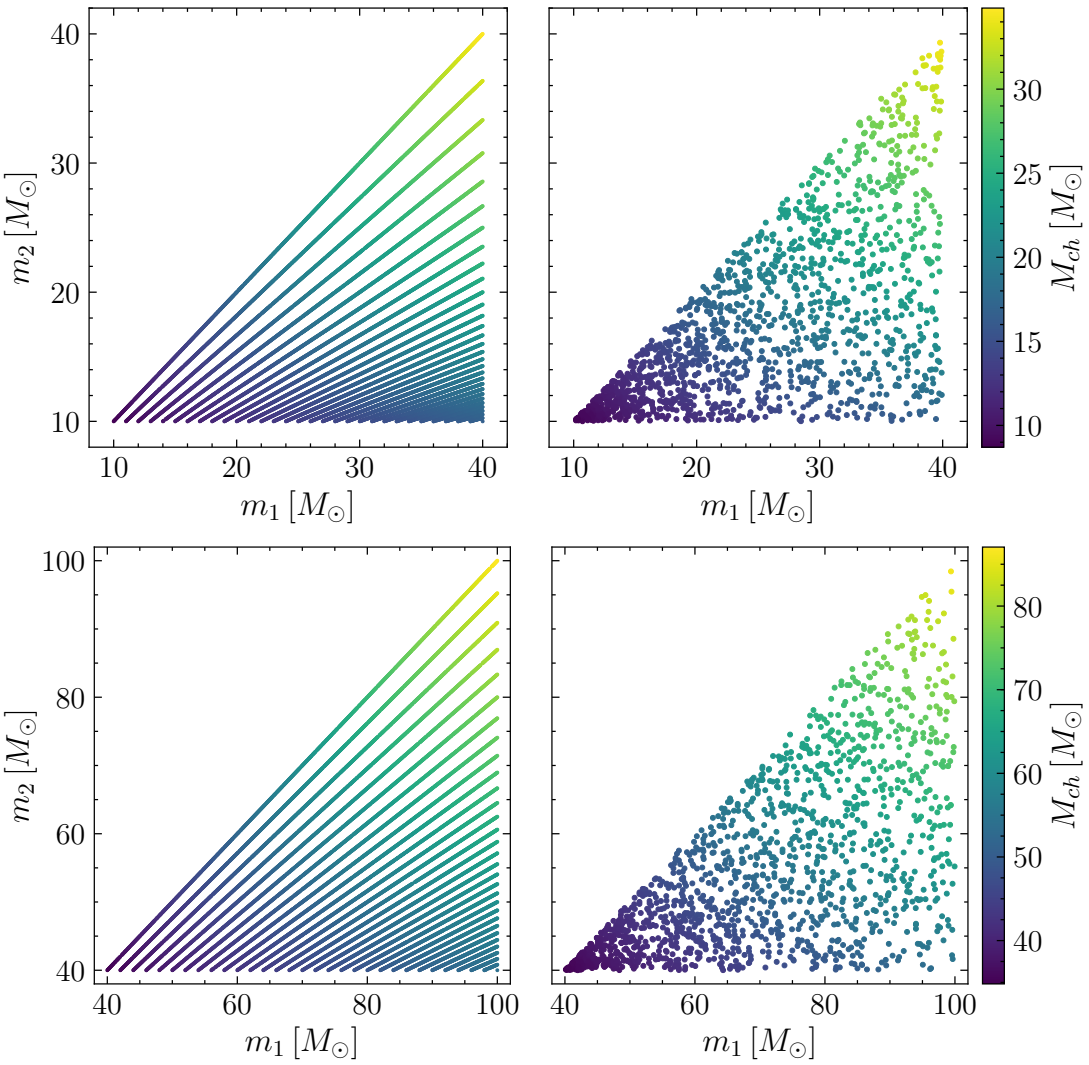


Figure 2.1 The component masses of training (left) and test (right) sub-datasets in dataset-1 (upper) and dataset-2 (bottom) with the color-coded chirp mass. While we use a set of fixed mass ratios, m_1/m_2, for the training sub-dataset, m_1 and m_2 are randomly chosen for the test sub-dataset with the restriction that $m_1 \geq m_2$. The mass ratios range from 1 to 4 for the dataset-1 and from 1 to 2.5 for dataset-2.

than using approximants in the sense of accuracy. However, we find that the number of publicly available NR-waveforms of BBHs is only $O(10^3)$ (Boyle et al. 2019; Healy et al. 2019; Jani et al. 2016; Healy et al. 2017). In specific, the number of non-spinning BBH waveforms reduces

to $O(10^2)$ (Boyle et al. 2019), so small that it might cause overfitting of the DL model (Groné 2017), which infects the general performance of the model. Thus we use EOB-waveforms to get a sufficient amount of training samples.

With the software and the approximant, we configure two datasets whose mass ranges of single black holes are $[10M_\odot, 40M_\odot]$ (dataset-1) and $[40M_\odot, 100M_\odot]$ (dataset-2) to divide search regions into low- and high-mass regions. Each dataset is consist of training, validation, and test sub-datasets with respective sample number ratio of 0.8, 0.1, and 0.1. The mass ratios of the sub-datasets are set differently[3]. For the training and validation samples, we use fixed mass ratios with an interval of 0.1 (0.05) within the range of [1, 4] ([1, 2.5]) for dataset-1 (dataset-2). On the other hand, we randomly sample m_1 and m_2 in the corresponding parameter space for the test sub-dataset. In this manner, we can prove that the model trained with a limited mass ratio samples can be applied to the ones residing in any regions of the parameter space. Fig. 2.1 shows the scatter plots of m_1 and m_2 of training sub-dataset in dataset-1 and -2 with color-coded chirp masses defined as $M_{ch} = (m_1 m_2)^{3/5}(m_1 + m_2)^{-1/5}$. We use the sampling rate, distance, and inclination angle of 4096Hz, 100Mpc, and 30°, respectively. The parameters employed for data preparation are tabulated in Table 2.1.

Following the data generation, the waveforms in dataset-1 and -2 are normalized with the maximum strain amplitude of each dataset. Since the diverse range of samples may cause biased results (Sola & Sevilla 1997), data normalization for the differently ranged dataset is necessary. By normalizing the dataset, the sample values can be restricted in a comparable range and contribute equally to the DL model optimization at the beginning of the training.

In turn, we divide each waveform into input and target waveforms: the input with the inspiral phase and target with merger and ringdown phases, respectively. For the division, we consider the point that the GW frequency reaches the innermost stable circular orbit (ISCO) frequency (Tredcr 1975) as the termination point of the inspiral phase (Favata 2011). The final data point of the input waveform and the initial data point of the target waveform are intentionally superposed to check whether the DL-waveform and given inspiral waveform are smoothly connected. Fig. 2.2 illustrates examples of input and target waveforms with different chirp

[3]The mass ratio is defined as m_1/m_2, and $m_1 \geq m_2$ is assumed by convention.

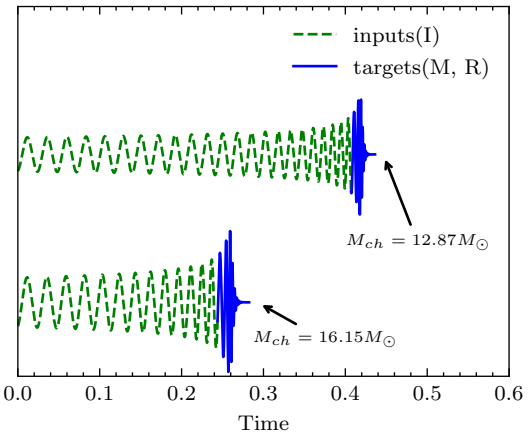


Figure 2.2 Examples of input (green dashed; inspiral) and target (blue solid; merger-ringdown) waveforms drawn with different chirp masses of the compact binary system. They are computed by using `SEOBNRv4`. The upper and lower waveforms depict $M_{ch} = 12.87M_{\odot}$ and $M_{ch} = 16.15M_{\odot}$, respectively. Note that the length of the generated waveforms changes depending on the mass.

masses. For the training of our DL model, we feed the input waveform to the DL model and let the model recover target waveform.

For divided target waveforms, we illustrate the number density distributions of waveform lengths in Fig. 2.3 (denoted by L_t). As shown in the figure, the distributions are not uniform. We reckon that this non-uniformity causes L_t-dependent accuracy of the DL model, which will be discussed in Sec 2.4.1.

2.3 Method

Since the duration of the GW emission within the detector's sensitive frequency band varies depending on the component masses or chirp mass of the binary system, we need a DL model capable of handling different size data. For this, we design a DL model with a novel architecture based on seq2seq, which is built for NLP. In this section, we briefly overview the original seq2seq model[4] and elaborate on our model below.

[4]For more details of the original model, we refer to (Cho et al. 2014; Sutskever et al. 2014).

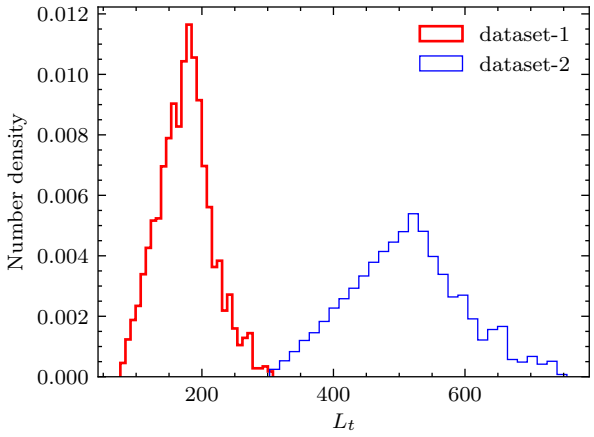


Figure 2.3 target waveform length (L_t) distribution of the training sub-dataset in dataset-1 (thick red) and dataset-2 (thin blue). Note that the non-uniform distributions are caused by the parameter sampling and input-target split method described in Sec. 2.2.

2.3.1 Original Sequence-to-Sequence Model

DL models for NLP take a batch of sentences as inputs and output transformed sentences. For that, each word in the sentences should be digitized since machine learning models work numerically. With the linguistic property that the number of vocabularies in a specific language is limited to a finite number, each distinct word can be represented as a vector by word embedding (Bengio et al. 2001). Thus, the sentence prediction problem can be regarded as selecting words from a given dictionary. The vectorized sentences, however, have different sizes because every sentence is composed of a different number of words.

To resolve the issue, the encoder, mapping the variable size input sequence into a fixed-size vector, is employed in the seq2seq model. Afterward, the transformed vectors, so-called representations, by the encoder are transmitted to the decoder, and it sequentially recovers the variable size target sentences. In the decoder calculation process, the output at the previous computing-step is taken as the input of the next step. Each sentence is required to end with the end-of-sequence token (⟨EOS⟩), and the decoder starts and finishes its computation by taking and outputting ⟨EOS⟩. The conditional vector ⟨EOS⟩ can be defined differently depending on the user's preference.

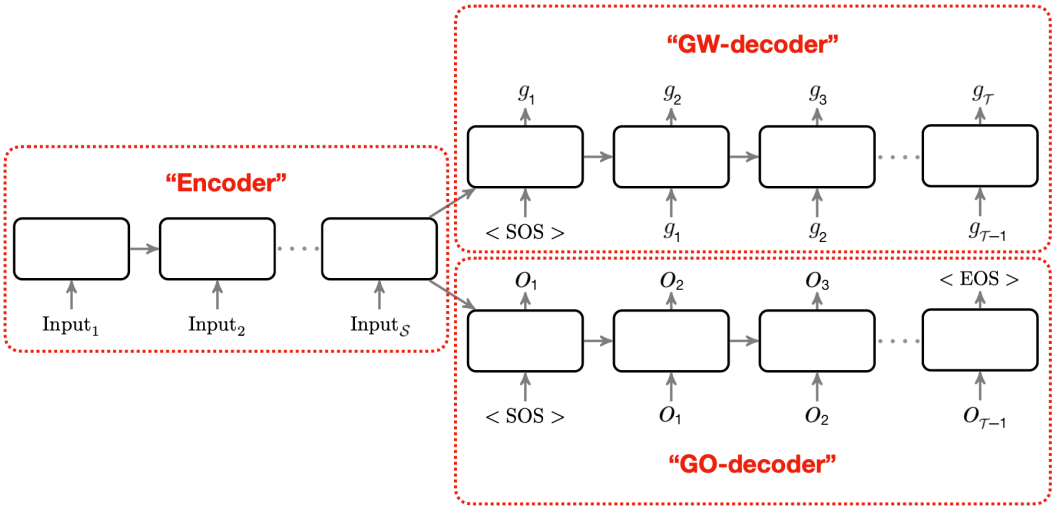


Figure 2.4 The schematic workflow of the DDS2S model. The solid black boxes indicate RNN cells. The model sequentially takes $\mathcal{S}$ vectors as input waveforms and attempts to regenerate target waveforms and GO-function, $\mathcal{G}$. The decoders start computation when inputted ⟨SOS⟩ and retrieve $\mathcal{T}$ vectors as output waveforms until the GO-decoder yields a value under 0.5, marked by ⟨EOS⟩. Note that the decoders use the output of the previous computing-step as the input at the next computing-step. The detailed structural information of the model is tabulated in Table. 2.2.

2.3.2 Dual-Decoder Sequence-to-Sequence Model

In the work of the original model, Sutskever et al. (Sutskever et al. 2014) were able to construct the ⟨EOS⟩, the interrupting condition of the decoder computation, using the linguistic property that the number of vocabularies is limited. Since the words in the dictionary can be discretely distinguished, it is clear to set the condition.

Regarding the GW-data, however, it becomes hazy to establish a criterion for interrupting the computing-step because the strain amplitudes of GWs are continuous real numbers: the number of possible cases is infinite, unlike the words in a dictionary. Thus, we cannot expect the model to produce an output that exactly matches a specific number by all digits. For example,

when we set $\langle \mathrm{EOS} \rangle = 0$, the model is unlikely to obtain the exactly matching value in machine precision.

As a strategy for learning this continuous sequence, we design a modified seq2seq model (DDS2S, Fig. 2.4) with one encoder and dual-decoder, GW- and GO-decoder: the encoder encrypts input waveforms, GW-decoder recovers target waveforms, and GO-decoder predicts the length of the target waveforms. While the computational mechanisms of the encoder and decoders are identical to the ones in the original model, the approach for handling input and target data is different.

First, the input and target waveforms are divided into the number of $\mathcal{S}$ and $\mathcal{T}$ vectors with $\mathcal{R}$ elements. When $\mathcal{R} > 1$, the ends of the waveform elements are zero-padded before division to match the component numbers with the multiples of $\mathcal{R}$. The zero-padded lengths of input and target waveforms can be computed via $L_s = \mathcal{S}\mathcal{R}$ and $L_t = \mathcal{T}\mathcal{R}$ [5]. Then, the encoder sequentially takes $\mathcal{R}$ elements of input waveforms $\mathcal{S}$ times and encrypts them into fixed-size vectors. The encoder outputs are transmitted to GW- and GO-decoders.

Subsequently, the GW-decoder regenerates $\mathcal{R}$ elements of target waveforms at every computing-step throughout the $\mathcal{T}$ step[6]. The generated waveforms are stacked in the order of computing-step and compared with the target waveforms to calculate the error function. As the error function of the GW-decoder, $\mathcal{I}$, we use the sum of mean-squared error and negative cosine similarity;

$$\mathcal{I}(g,t) = \frac{1}{\mathcal{T}}\Sigma_i (g_i - t_i)^2 - \frac{g \cdot t}{||g||\,||t||}, \tag{2.1}$$

where g and t are respectively the generated and target waveforms; $\mathcal{T}$ is the number of vectors for the given target waveform; $||\cdot||$ is L^2 norm.

Lastly, we establish the GO-function to endow the GO-decoder the capability to estimate the length of the target waveform. When the given target waveform consists of $\mathcal{T}$ vectors, we can set the integer condition, C, for progressing from computing-step τ to $\tau + 1$ as follows: 1 for proceeding and 0 for breaking.

[5] Note that L_s and L_t are the lengths of input and target waveforms without zero-padding as $\mathcal{R} = 1$.

[6] The total computing-step multiplied by $\mathcal{R}$ and waveform length are compatible concepts, and one can convert them into the duration of GW by multiplying the inverse of the sampling rate, 4096Hz.

$$C_\tau = \begin{cases} 1, & \text{if} \quad 1 \le \tau < \mathcal{T} - 1 \\ 0, & \text{if} \quad \tau \ge \mathcal{T}. \end{cases} \tag{2.2}$$

We may use the set of C_τ to train GO-decoder, but the discrete values and rapid decrease of C from $\tau = \mathcal{T} - 1$ to $\tau = \mathcal{T}$ are inappropriate for the training of the DL model. Thereby, we define GO-function, $\mathcal{G}$, approximating the integer C values with a smooth decreasing pattern near $\tau = \mathcal{T}$ and use the function to compute the mean-squared error with the GO-decoder outputs. The GO-function and the error function, $\mathcal{J}$, of the GO-decoder are described below.

$$\mathcal{G}_\tau = \begin{cases} 1 - 0.5\,(\tau/\mathcal{T})^\alpha, & \text{if} \quad 1 \le \tau \le \mathcal{T} - 1 \\ 0, & \text{if} \quad \tau \ge \mathcal{T}, \end{cases} \tag{2.3}$$

$$\mathcal{J}(o, \mathcal{G}) = \frac{1}{\mathcal{T}} \Sigma_i (o_i - \mathcal{G}_i)^2, \tag{2.4}$$

where o_i is the output of GO-decoder. Fig. 2.5 presents how the curve of the $\mathcal{G}$ varies according to different αs. As the α is getting bigger, the GO-function approximates the C values more accurately. On the contrary, we find that the rapid decrease of $\mathcal{G}$ near $\tau = \mathcal{T}$ hinders the training of the DL model when the α is too high. We empirically determine α of 5 for the training of the model.

The final loss for the training is the sum of the error function of GW- and GO-decoders, namely $\mathcal{I} + \mathcal{J}$. The model is trained by adjusting its parameters in such a way the error is minimized.

We apply the `Sigmoid` to the output layer of the GO-decoder since the GO-function should output values from 0 to 1. Then, we have given output values rounded to either 0 or 1. The computation continues when the rounded value is 1 and stops otherwise. Hence, the GO-decoder output below 0.5 serves as ⟨EOS⟩ in our case. For this reason, we define $\mathcal{G}$ to have a slightly higher value than 0.5 at $\tau = \mathcal{T} - 1$ because we expect the model to stop calculating at $\tau = \mathcal{T}$. For the DDS2S model, we newly define zero vectors with $\mathcal{R}$ elements as a start-of-sequence token (⟨SOS⟩), which is inputted at the start of decoder computation.

Among the prominent RNN cells, we choose Gated Recurrent Unit (GRU) (Cho et al. 2014) for the encoder and both decoders because the setting with GRU showed higher accuracy

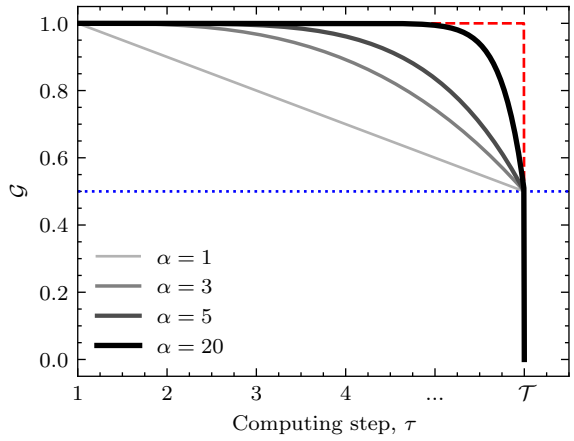


Figure 2.5 The GO-function, $\mathcal{G}$, with several values of α in greyscale. The red dashed-line depicts how the integer condition, C, changes according to the computing-step. As the value of the α increases, the function approximates the C values more accurately. We also draw the horizontal blue dotted-line at 0.5, the condition of interrupting decoders' computation.

Table 2.2 Detailed structure of the DDS2S model.

	Encoder	GW-Decoder	GO-Decoder
RNN cells	GRU	GRU	GRU
The number of RNN cells	$\mathcal{S}$	$\mathcal{T}$	$\mathcal{T}$
The number of input layers	1	1	1
The number of hidden layers	4	4	4
The number of output layers	-	1	1
The number of input neurons	$\mathcal{R}$	$\mathcal{R}$	1
The number of hidden neurons	256	256	256
The number of output neurons	-	$\mathcal{R}$	1
Activation function of input layers	`Tanh`	`Tanh`	`Tanh`
Activation function of hidden layers	`Tanh`	`Tanh`	`Tanh`
Activation function of output layers	-	-	`Sigmoid`

and faster training than Long-Short Term Memory (Hochreiter & Schmidhuber 1997; Gers et al. 1999), another well-known RNN cell. A fully connected layer is placed at the end of the

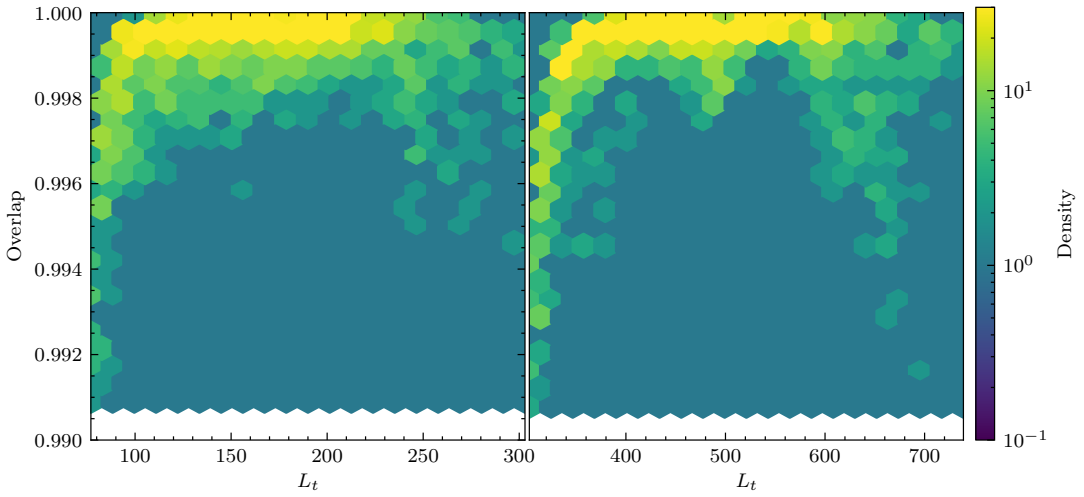
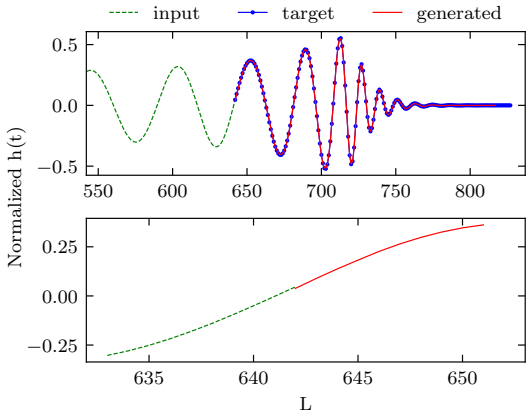


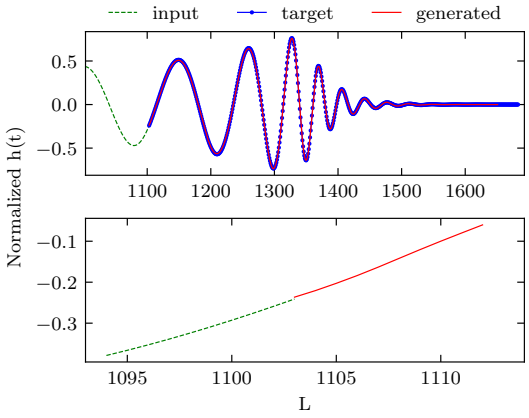
Figure 2.6 Density heatmap of overlap according to target waveform lengths, L_t, for the dataset-1 (left) and 2 (right). We draw the vertical axes of the two plots in the same range and scale. For clear contrast, we leave the regions with no samples empty at the bottom of the plots. As shown in the plots, overlaps of all the DL-waveforms are higher than 0.990. Besides, the averages of the waveforms from both datasets are over 0.999. However, a few shortest and longest samples have smaller overlap values. Considering the relatively small number of the shortest and longest waveforms in the training sub-dataset (Fig. 2.3), it implies that the non-uniformity of the sub-dataset is related to the locally different accuracy of the DL model.

decoders' hidden layers to convert hidden states to vectorized outputs with $\mathcal{R}$ components. We use the hyperbolic tangent as the activation function for hidden layers of each RNN cell of encoder and decoders.

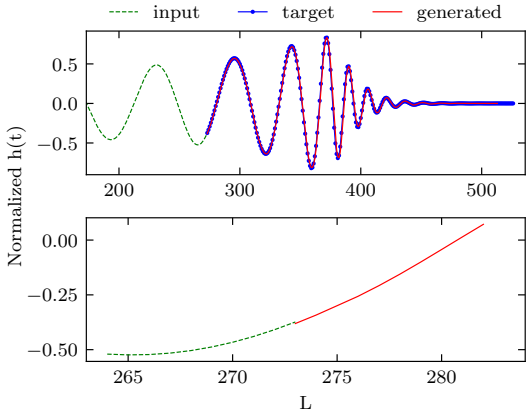
For the model structure, we find an empirically optimal model configuration varying the number of neurons in hidden layers (hereafter, hidden neurons) based on the overlap to a reference waveform, which we will discuss in the following sub-section. The information on the network configurations and hyperparameters of the optimal model is summarized in Table 2.2.



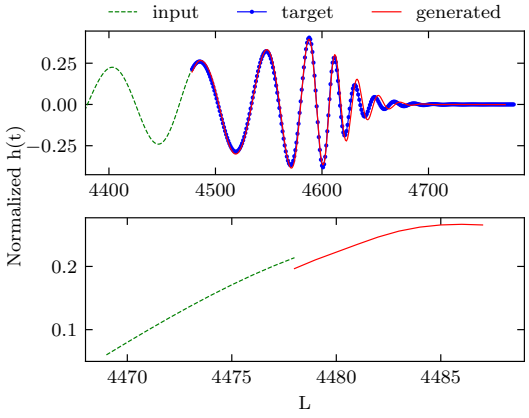
(a) Best case of dataset-1, $\mathcal{M} = 0.999$



(b) Best case of dataset-2, $\mathcal{M} = 0.999$



(c) Worst case of dataset-1, $\mathcal{M} = 0.991$



(d) Worst case of dataset-2, $\mathcal{M} = 0.991$

Figure 2.7 The input (green dashed), target (blue solid line with dots), and DL- (red solid) waveforms from dataset-1 (left column) and dataset-2 (right column) with the amplified images of connection points. The horizontal and vertical axes indicate the length of the waveforms in sampling unit and the normalized strain amplitude of the GWs, respectively. We only show a hundred sampling units of input waveforms in the plots for clear visualization. The top and bottom panels are the waveforms with the highest and lowest overlap cases, respectively.

2.3.3 Overlap

We use overlap to assess the DL-waveforms' accuracy. The normalized overlap, $\mathcal{M}$, of the DL-waveform g and the target t can be computed as

$$\mathcal{M} \equiv \frac{(g|t)}{\sqrt{(g|g)(t|t)}}, \tag{2.5}$$

where $(g|t) = \int_{-\infty}^{\infty} \tilde{g}(f)\tilde{t}^{*}(f)df$. $\tilde{g}$ and $\tilde{t}$ are the Fourier transform of g and t, respectively, and asterisk mark (*) is complex conjugate. Note that $\mathcal{M}$ becomes 1 for the perfect match and 0 for the perfect mismatch between g and t.

From the grid-search described in Appendix 2.A, we choose an empirically optimal model configuration, maximizing the minimum overlap of the model's output waveforms. Providing accuracy, we use the setup with 256 hidden neurons and $\mathcal{R} = 1$. Henceforward, we shall only discuss the results of the model with 256 hidden neurons and $\mathcal{R} = 1$. The detailed explanation can be found in Appendices 2.A and 2.B.

2.4 Result

2.4.1 Waveform Validation

The Fig. 2.6 depicts the overlap density heatmap between the DL-waveforms and corresponding target EOB-waveforms of the dataset-1 and 2. All of the DL-waveforms are in excellent agreement with their target waveforms in both cases as the minimum value of overlaps is higher than 0.990[7]. Furthermore, the mean overlaps of waveforms from both datasets are higher than 0.999, indicating less than 0.1% average error.

However, as we can see from the figure there are several outliers whose overlaps are substantially smaller than the majority. We explore the dependence of the overlap on the target waveform length to track down possible reasons for relatively poor overlap cases. The heatmap shows the distribution of the overlaps concerning the length of the target waveforms. The overlap of dataset-1 (dataset-2) tends to decrease at the short-end and long-end of the target waveform length, i.e., $L_{\mathrm{t}} \lesssim 100$ or $L_{\mathrm{t}} \gtrsim 250$ ($L_{\mathrm{t}} \lesssim 400$ or $L_{\mathrm{t}} \gtrsim 600$). As shown in Fig. 2.3, the training samples in the range of $100 \lesssim L_{\mathrm{t}} \lesssim 250$ of dataset-1 and $400 \lesssim L_{\mathrm{t}} \lesssim 600$ of dataset-2 dominate the number distribution of the target waveform length. It can be attributed to the fact that the model is more likely to weigh the majority of the training sub-dataset.

[7]For comparison, the authors of Ref. (Sturani et al. 2010) have shown that the overlap between numerical and their phenomenological waveforms ranges from 0.95 to 0.99. On the other hand, Ref. (Wei & Huerta 2020) have shown their model results in the overlap ≥ 0.99.

We also visually inspect the agreement between the DL-waveforms and target waveforms. Fig. 2.7 shows the best and worst overlap cases of the DL-waveforms. The overlaps of the best cases for both datasets are $\mathcal{M} = 0.999$. The time-series of the DL-waveforms matches well with the target waveforms. For the worst cases, the overlaps of the two datasets are both 0.991 (Fig. 2.7 (c) and (d)). We see that there exist small discontinuities between the DL- and input waveforms as shown in the lower panel of the figure. We may resolve the discontinuity by post-processing or letting the DL model generate the whole waveform in the inspiral-merger-ringdown phase at once. We leave this issue to future work.

2.4.2 Injection Test

Next, we attempt to use the DL-waveform templates in simplistic search of parameters, i.e., m_1, m_2 and the event time, of the simulated GW signals. To replicate practically used waveform templates, we hybridize inspiral `SEOBNRv4`-waveform and merger-ringdown DL-waveform by simply concatenating the two waveforms. One may implement sophisticate hybridization of waveforms, but it is beyond the scope of this work. We perform parameter grid-search instead of Markov Chain Monte Carlo, typically executed for the parameter estimation of GWs (Van Der Sluys et al. 2008), due to the practical difficulty of plugging a new waveform model in the existing parameter estimation code (Aasi et al. 2013). For the computation of SNR and the search of the events, the matched filtering engine of `PyCBC` (Usman et al. 2016) is used.

To simulate the observation data embracing a GW signal, we use the LIGO-Hanford O1 data provided by GW Open Science Center[8]. We randomly select a 32-second segment from the data without any known GW signals and inject a `SEOBNRv4`-waveform into the center. While we use five sets of different injection parameters and distances, we fix the inclination angle to 30°for simplicity. The configuration setups of the tests are tabulated in the first three columns of Table 2.3.

By performing the parameter grid-search for multiple injection waveforms, we retrieve injection parameters in all examinations within the 90% confidence interval. We first define the search parameter sets, (m_1, m_2) on regularly-spaced grid of the parameter space. Then,

[8]https://www.gw-openscience.org/archive/O1/

we construct the full IMR waveform templates by hybridizing the inspiral waveform and the merger-ringdown DL-waveform using `SEOBNRv4` and DDS2S, respectively, for the parameter sets. Across the parameter sets, we compute SNR by matched filtering with each waveform template using `PyCBC` on the simulated data. Assuming the likelihoods of the parameter sets are proportional to the SNR, we estimate the probability density function (PDF) of the parameters. Then, we marginalize the PDF with respect to each parameter and acquire the median as the best-fit parameters with their 90% confidence interval. Subsequently, we repeat the entire process with different combinations of injection masses and distances. The best-fit parameters with confidence intervals and their SNRs are summarized in the last two columns of Table 2.3.

The best-fit parameters and the high SNR region emerge around the chirp mass contour line of the injected signal. Since the chirp mass of GW is governed by the frequency and frequency derivative (Abbott et al. 2017), and its SNR depends on frequency evolution (Flanagan & Hughes 1998), the SNR of GW again relies on the chirp mass. It is well-reflected in the example contour map of the signal with $m_1 = 25M_\odot$ and $m_2 = 15M_\odot$ (Fig. 2.8).

Using the best-fit parameters found from the grid search, we perform event time searches and find the SNR peak at where we inject the signals. We illustrate SNR time-series of the above example case in Fig. 2.9. As can be seen in the figure, the peak SNR occurs at the center of the data segment, where we have injected the simulated signal.

It is known that the systematic error from waveform approximants is independent of SNR, while the statistical error due to noise roughly scales as $1/\mathrm{SNR}$. One can readily expect that the systematic error could dominate in higher SNR signals. Cutler and Vallisneri Cutler & Vallisneri (2007) have presented rigorous computation of the systematic errors in parameter estimation using 3.5PN (post-Newtonian approximation of order 3.5) waveforms for inspiral signals of massive black hole binaries. They have shown that the magnitude of the systematic errors from 3.5PN waveforms with $\mathcal{M} > 0.9999$ commensurate with the SNR ~ 1000 statistical errors. Motivated by this, we roughly estimate the impact of systematic error of our DL-based waveform on the parameter estimation by repeating the grid-search of parameters as described above with varying SNR of the injected signal. By comparing the systematic error with the statistical errors of the same parameter as increasing the SNR of the injected signal, we find

Table 2.3 Summarized results of the injection tests. The best-fit parameters and their SNR for the injected signals are computed by `PyCBC` matched filtering engine with DL waveform templates. We establish template waveforms by hybridizing inspiral `SEOBNRv4` and merger-ringdown DL waveforms. The m_1 and m_2 are given in the unit of the solar mass. I, M, and R indicate inspiral, merger, and ringdown phases, respectively.

Template approximant	Distance (Gpc)	Injection (m_1, m_2)	Best-fit (m_1, m_2)	SNR
EOB (I) + DL (MR)	1.6	80.0, 65.0	$80.1^{+13.7}_{-14.5}$, $61.7^{+18.3}_{-16.4}$	14.5
	1.5	70.0, 60.0	$73.9^{+16.5}_{-16.9}$, $58.6^{+16.6}_{-14.4}$	13.0
	0.8	35.0, 20.0	$33.1^{+5.6}_{-6.8}$, $21.5^{+9.0}_{-8.4}$	12.7
	0.7	30.0, 25.0	$31.6^{+6.3}_{-7.0}$, $22.7^{+8.6}_{-8.3}$	15.3
	0.6	25.0, 20.0	$28.3^{+8.0}_{-8.4}$, $18.9^{+7.1}_{-6.6}$	15.7

that the magnitude of the systematic error becomes comparable to the 1-σ statistical error at SNR $\sim O(10)$ in our DL-based waveform approximant.[9]

2.4.3 Performance Dependence on the Dataset Size

We inspect the dependence between the accuracy of the DL model and the number of waveforms in the training sub-dataset. The test is performed to explore the viability of applying the proposed model to NR-waveforms, in which only a few thousands exist (Boyle et al. 2019; Healy et al. 2019; Jani et al. 2016; Healy et al. 2017). We generate four reduced datasets with half and the tenth number of waveforms in the original training data of dataset-1 and -2, maintaining the number of waveforms in the validation and test data.

We find that one-tenth of the original size is enough to reach the required accuracy of $\mathcal{M} \geq 0.99$. The model is trained more than five times with each reduced training data. It turns out that the minimum and average values of overlap are higher than 0.990 and 0.999, equivalent to 1.0% and 0.1% error, respectively, for all DL-waveforms of the trained model from each run. The mean values for the averaged overlaps and minimum overlaps from more

[9]Note that our approach for finding the SNR level where the two errors become similar is not rigorous. For a more in-depth exploration of the systematic errors, refer to Cutler & Vallisneri (2007).

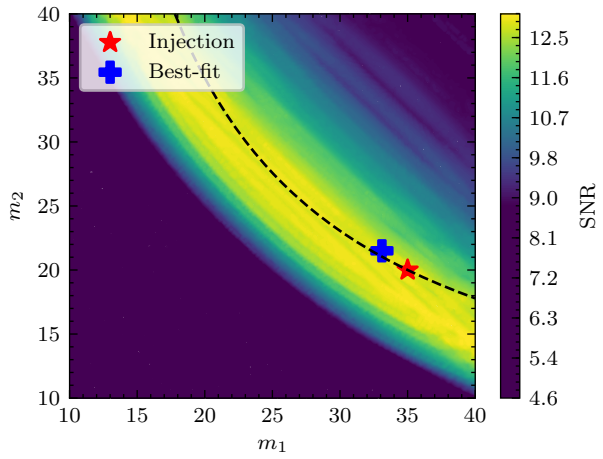


Figure 2.8 Filled contour map of SNR in the parameter space for the injection signal with $m_1 = 35M_\odot$ and $m_2 = 20M_\odot$. Each of the red star and blue plus markers indicates injection and best-fit parameters. The black dashed line is a contour with the level of injection chirp mass. The best-fit parameters and the high SNR region arise in the vicinity of the contour line. Although our parameter space is restricted with the condition $m_1 \geq m_2$, the filled contour map is reflected on the slope of 1 line for aesthetic visualization.

than five individual runs are tabulated in Table 2.4. We also present the results of Sec 2.4.1 for comparison in the last column. The relative dataset size in the table means the ratio of the number of waveforms in the training data to the number of waveforms in the original training data. The result shows that reducing the number of waveforms down to 1000 for the training hardly affects obtaining the desired accuracy. Hence, we advocate that the application of the DDS2S model to NR-waveforms is feasible.

2.5 Summary and Discussion

The efficiency of matched filtering for searching GW signals buried in noisy GW data has been proved by recent successful detections of GW signals. Although NR can increase the accuracy of template waveforms, expensive computational costs of running NR limit the use of it for the generation of a sufficiently large number of template waveforms. This drawback

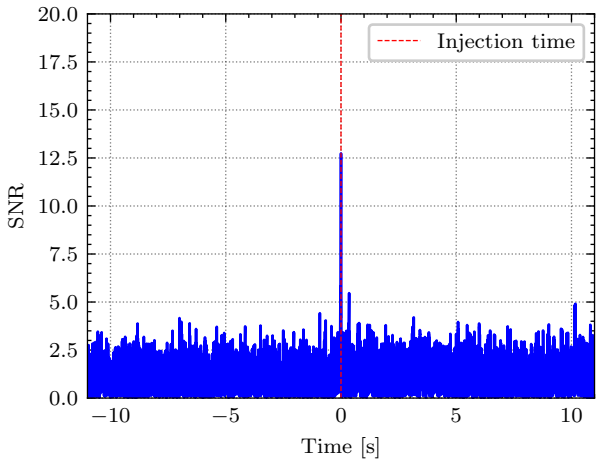


Figure 2.9 SNR time-series computed by matched filtering engine of `PyCBC` and best-fit DL-waveform template of Fig. 2.8. The injected signal is the `SEOBNRv4`-waveform ($m_1 = 35M_\odot$ and $m_2 = 20M_\odot$). Here, we initialize the start time of the injected signal to 0, marked by the red dashed line. Note that the SNR peak occurs at the injection time.

Table 2.4 Accuracy variation of the DL model according to dataset size. We also show the results of the Sec. 2.4.1 in the last column of the table for comparison. The mean values for the minimum and average overlaps from more than five individual runs for each dataset are summarized in the table. The value of the relative dataset size is the ratio of the number of waveforms between the reduced training sub-dataset and the full sub-dataset introduced in Sec. 2.2.

Relative dataset size	0.1	0.5	1
Minimum overlap (dataset-1)	0.991	0.990	0.991
Minimum overlap (dataset-2)	0.990	0.990	0.991
Average overlap (dataset-1)	0.999	0.999	0.999
Average overlap (dataset-2)	0.999	0.999	0.999

of NR eventually led to the use of approximate waveforms for the matched filtering instead. Motivated by such difficulties, we have examined the DL method for the generation of template waveforms with much smaller computational costs but comparable accuracy to NR.

To study the feasibility of this consideration, we have implemented the DDS2S model. The encoder-decoder structure is capable of handling the variable sizes of different waveforms, and the dual-decoder structure enables the model to control the continuous real-numbered sequences.

We also have examined the applicability of the waveforms by computing the overlap with EOBNR-based waveforms and performing the injection test. The accuracy of the DL-based waveforms is found to be better than 99.9 % in most combinations of the masses, while a small number of outliers with overlap as small as 0.99 exists. In the injection test, we have recovered the event time of waveforms injected into real noise data with the conventional matched filtering engine of `PyCBC`.

We have found that the method generating merger-ringdown waveforms using the inspiral waveforms needs to be improved. For example, we have seen that discontinuities occurred between input and output waveforms, as shown in Fig. 2.7, although the minimum overlap of DL-waveforms to the EOB-waveform was higher than 0.990. To avoid this issue, we may take a new strategy of generating a full IMR waveform. However, the main goal of this chapter is to demonstrate the feasibility of adopting DL to model the merger-ringdown waveforms. We leave the implementation of a DL model generating the full waveforms to future work.

Regarding the speed of waveform generation, the DDS2S model has an advantage over other waveform approximants when computing a batch of multiple waveforms simultaneously. For computing a single waveform, EOB is faster than the DDS2S model, typically taking $O(10^{-2})$ seconds using a modern CPU core. However, the DDS2S model generates $\sim$ 1500 waveforms using pre-generated inspiral waveforms in $O(1)$ seconds using NVIDIA GeForce GTX 1080, while EOB took $O(10)$ seconds. The disparity arises since the DL models are specialized for batch computations, which process multiple data at once.

The DDS2S model has been built to learn how to predict the output waveforms only from the given input waveforms without any specific physical information of the source binary system. Thus, we can readily extend this work to various systems of interest.

For a more precise description of realistic physical binary systems, we need to have waveform models for more complex binaries: a wider range of the mass ratios, the spin of each com-

ponent, eccentricity of the orbits. GWs from unbound orbit such as hyperbolic and parabolic encounters are also of great interest. Lastly, it is worthwhile to mention that recalibration of full IMR waveforms to increased amounts of NR waveform data is in progress in the community. (The LSC-Virgo-KAGRA Observational Science Working Groups 2020)

Our approach described in this chapter can potentially be applied to more complex systems described above because DDS2S only depends on training data, not any assumptions or approximations on which other waveform models are based. Moreover, we have observed that $\sim$ 1000 training waveforms are sufficient for the model to reach the expected level of accuracy in Sec 2.4.3. Thus, as long as there is a sufficiently large number of training waveform samples for any systems or NR are given, DDS2S can be trained to generate accurate waveforms in principle.

2.A Empirically Optimal Number of Hidden Neurons

We investigate the influence of the hidden neurons on the accuracy of the models; 64, 128, and 256 hidden neurons.

Accuracy-wisely, we find that the model with 256 hidden neurons is most suitable amid the tested cases. To compare model accuracy according to the number of hidden neurons, minimum and average overlap between DL-waveforms and corresponding target waveforms are computed. Table 2.A.1 summarizes the minimum and average overlaps of the models for dataset-1 and -2. The minimum overlap values of each model from dataset-1 (dataset-2) are 0.984, 0.990, and 0.991 (0.977, 0.989, and 0.991) in the increasing order of the model size. All of the average overlaps are the same as 0.999, except the case of the smallest model with dataset-2, whose overlap is 0.998 (overlaps of 0.999 and 0.998 are equivalent to 0.1% and 0.2% errors). Namely, the model with 256 hidden neurons shows the highest accuracy.

Table 2.A.1 Minimum and average overlap values of the test sub-dataset in dataset-1 and -2 according to models with the different number of hidden neurons.

The number of hidden neurons	64	128	256
Minimum overlap (dataset-1)	0.984	0.990	0.991
Minimum overlap (dataset-2)	0.977	0.989	0.991
Average overlap (dataset-1)	0.999	0.999	0.999
Average overlap (dataset-2)	0.998	0.999	0.999

2.B Computing Time and Accuracy Variation of The Model According To $\mathcal{R}$

We examine how the number of elements $\mathcal{R}$ in an RNN cell affects the model in the aspects of computing time and accuracy. Table 2.B.2 tabulates the typical elapsed time with a minimum overlap of each case on dataset-1 and -2. Although the model can speed up by increasing $\mathcal{R}$, the accuracy expense renders the model inapplicable for practical use.

Table 2.B.2 Computation time and overlap variation with respect to the number of elements, $\mathcal{R}$, in a RNN cell.

$\mathcal{R}$	T_1	T_{1500}	Minimum overlap	
			dataset-1	dataset-2
1	$O(10^{-1})$	$O(1)$	0.991	0.991
10	$O(10^{-2})$	$O(10^{-1})$	0.913	0.910
100	$O(10^{-3})$	$O(10^{-2})$	0.823	0.805

Chapter 3

Estimation of Photometric Redshifts. I. Machine Learning Inference for Pan-STARRS1 Galaxies Using Neural Networks[1]

Abstract

We present a new machine learning model for estimating photometric redshifts with improved accuracy for galaxies in Pan-STARRS1 data release 1. Depending on the estimation range of redshifts, this model based on neural networks can handle the difficulty for inferring photometric redshifts. Moreover, to reduce bias induced by the new model's ability to deal with estimation difficulty, it exploits the power of ensemble learning. We extensively examine the mapping between input features and target redshift spaces to which the model is validly applicable to discover the strength and weaknesses of trained model. Because our trained model is well calibrated, our model produces reliable confidence information about objects with non-catastrophic estimation. While our model is highly accurate for most test examples residing in

[1] Submitted to the Astronomical Journal

the input space, where training samples are densely populated, its accuracy quickly diminishes for sparse samples and unobserved objects (i.e., unseen samples) in training. We report that out-of-distribution (OOD) samples for our model contain both physically OOD objects (i.e., stars and quasars) and galaxies with observed properties not represented by training data. The code for our model is available at `https://github.com/GooLee0123/MBRNN` for other uses of the model and retraining the model with different data.

3.1 Introduction

For various astronomical studies, the photometric redshifts of galaxies are critical. Representative research areas include cosmological model testing (Blake & Bridle 2005; Amon et al. 2018) and dark energy survey (Banerji et al. 2008; Sanchez et al. 2014). In terms of accuracy, although the spectroscopic estimation of redshifts is the most appropriate method, acquiring spectroscopic redshifts is significantly more expensive than estimating photometric redshifts (Salvato et al. 2019). In terms of cost, at a tolerable expense of accuracy, photometric redshifts can be a suitable substitute for spectroscopic redshifts.

Modern photometric redshift estimation approaches are split into two large branches: the spectral energy distribution (SED) fitting based on SED models (including spectral templates) and machine learning inference (Cavuoti et al. 2017; Salvato et al. 2019). These two methods are mutually complementary with different pros and cons. The template-based SED fitting may provide photometric redshifts in a wide redshift and photometric range; moreover, using Bayesian inference improves the effectiveness of the method (Bolzonella et al. 2000). However, this approach heavily relies on the prior knowledge of SEDs and the understanding of related physics determining SEDs. This dependency may result in biased results (Walcher et al. 2011; Tanaka 2015). However, the machine learning method can quickly retrieve accurate photometric redshifts without dependence on prior knowledge (Cavuoti et al. 2017). Nonetheless, most of the machine learning models suffer from performance degradation for few or unseen data during their training since these methods are induction models for the provided data (Liang et al. 2017; Hendrycks et al. 2019).

Samples drawn from out-of-distribution (OOD) (i.e., few or unseen data in training) are

well-known distress in a reliable application of neural networks (NNs). Hendrycks & Gimpel (2017) demonstrate that an ML model's accuracy degrades for OOD samples for several training datasets: e.g., MNIST (LeCun & Cortes 2010), CIFAR-10, and CIFAR-100 (Krizhevsky 2012), and they suggest a baseline model for OOD detection in an NN. K. Lee et al. (2018) propose an advanced method for detecting OOD examples using Mahalanobis distance, thus assuming the trained network parameters can be fitted by a class-conditional Gaussian distribution. The unsupervised approach introduced by Yu & Aizawa (2019) uses unlabeled samples as training data to equip NNs with the functionality of scoring and detecting OOD examples. A parameter-free OOD score is proposed by Serrà et al. (2020) to handle the OOD issue in generative models, thus posing that the problem is attributed to the excessive effect of input complexity. These recent studies emphasize that, for achieving a more robust NN model, the OOD instance is a practical and important limitation.

The appropriate warning on OOD samples and handling their impact on models should be offered in machine learning inference of photometric redshifts. The machine learning method for photometric redshift estimation has been explored in many past studies (Firth et al. 2003; Ball et al. 2008; Singal et al. 2011; Brescia et al. 2013; Laigle et al. 2017; Bilicki et al. 2018; Chong & Yang 2019). In these past studies, although the superior performance of machine learning methods has been demonstrated, the input/target feature spaces regarding OOD examples that the models are unable to describe for accurate and reliable prediction of photometric redshifts have not been quantitatively investigated.

In this series of studies, we propose a machine learning method to improve the accuracy and reliability of photometric redshift inference, thus exploiting the well-known flexibility of NN models. The NNs are renowned for their capacity of mapping nonlinear relationships between input and target (i.e., redshift) as a universal approximator and handling a massive amount of data. NNs have been one of the most popularly used machine learning algorithms throughout a wide range of fields and tasks including natural language processing (Vaswani et al. 2017; Brown et al. 2020), image classification (Krizhevsky et al. 2012; Szegedy et al. 2015), autonomous vehicles (Levinson et al. 2011; Dosovitskiy et al. 2017b), and protein structure prediction (Senior et al. 2020; Torrisi et al. 2020).

This study, the first study in the series, focuses on improving accuracy in inferring photometric redshifts. In our NN models, we adopt anchor loss (Ryou et al. 2019), which considers the difficulty of inferring photometric redshifts with respect to the estimated redshift, i.e., target. The primary cause of the inference difficulty is imbalanced training samples for redshifts and complex patterns in mapping from input features to the target redshift space. Because a new loss can cause systematic bias effects, we use an ensemble learning approach, which combines multiple base models into a unified model, to reduce the bias of models and improve accuracy (Zhou 2009).

Photometric data used are collected from the Panoramic Survey Telescope and Rapid Response System (Pan-STARRS) public data release (Chambers et al. 2016). We intend to use our trained model to infer the photometric redshifts of objects that correspond to extragalactic transient and variable sources. For transient sources such as supernovae and variable sources such as quasars, Pan-STARRS photometric catalogs can be useful to infer the photometric redshifts of hosts. Because photometric information used by our model is not much different from that acquired in other surveys, such as the SkyMapper Southern Sky Survey (Keller et al. 2007) and Legacy Survey of Space and Time (Tyson 2002; Ivezić et al. 2019), using the Pan-STARRS data for the model has potential benefits for applications with other data. Moreover, there are not many previous studies on photometric redshifts with Pan-STARRS data (Beck et al. 2021).

This chapter is organized as follows. Section 4.1 provides the overview of training data and pre-processing of input features for machine learning applications. In Section 4.2, machine learning approaches and performance evaluation metrics are elaborated in detail. Section 4.3 focuses on the performance analysis of our machine learning model and its comparison with baseline models. In Section 3.5, we explore the mapping between the input feature space and the target redshift space for which our model is validly applicable using comparison data. Finally, we discuss our results and provide the conclusion in Section 4.4.

3.2 Data

Training samples comprise 1,480,262 galaxy objects with known spectroscopic redshifts. We compile the samples from multiple spectroscopic redshift catalogs with the condition of reliable

Table 3.1 Spectroscopic galaxy redshift samples.

Dataset name	Number of objects	Selection conditions	Reference
SDSS DR15	1294042	(CLASS == GALAXY) and (ZWARNING == 0 or 16) and (Z_ERR >= 0.0)	Aguado et al. (2019)
LAMOST DR5	116186	(CLASS == GALAXY) and (Z > -9000)	Cui et al. (2012)
6dFGS	45036	(QUALITY_CODE == 4) and (REDSHIFT <= 1.0)	D. H. Jones et al. (2009)
PRIMUS	11012	(CLASS == GALAXY) and (ZQUALITY == 4)	Cool et al. (2013)
2dFGRS	7000	(Q_Z >= 4) and (O_Z_EM < 1) and (Z < 1)	Colless et al. (2001)
OzDES	2159	(TYPES ! = RadioGalaxy or AGN or QSO or Tertiary) and (FLAG ! = 3 and 6) and (Z > 0.0001)	Childress et al. (2017)
VIPERS	1680	(4 <= ZFLG < 5) or (24 <= ZFLG < 25)	Scodeggio et al. (2018)
COSMOS-Z-COSMOS	985	((4 <= CC < 5) or (24 <= CC < 25)) and (REDSHIFT >= 0.0002)	Lilly et al. (2007, 2009)
VVDS	829	ZFLAGS == 4 or 24	Le Fèvre et al. (2013)
DEEP2	540	(ZBEST > 0.001) and (ZERR > 0.0) and (ZQUALITY == 4) and (CLASS == GALAXY)	Newman et al. (2013)
COSMOS-DEIMOS	517	(REMARKS ! = STAR) and (QF < 10) and (Q >= 1.6)	Hasinger et al. (2018)
COMOS-Magellan	183	(CLASS == nl or a or nla) and (Z_CONF == 4)	Trump et al. (2009)
C3R2-Keck	88	(REDSHIFT >= 0.001) and (REDSHIFT_QUALITY == 4)	Masters et al. (2017, 2019)
MUSE-Wide	3	No filtering conditions.	Urrutia et al. (2019)
UVUDF	2	Spectroscopic samples.	Rafelski et al. (2015)

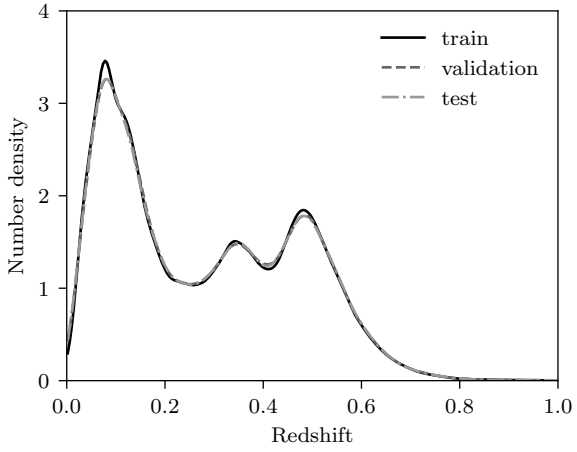


Figure 3.1 Redshift distributions of training, validation, and test sets. We use an adaptive kernel density estimation for density estimation of redshifts.

redshift estimation. Table 3.1 summarizes the spectroscopic redshift samples that satisfy the required selection conditions and have acceptable photometric data as subsequently described. Most selection conditions adopted here help us use only samples with highly reliable redshifts.

We use the photometric data of Pan-STARRS project as input photometric data for the machine learning model. The Pan-STARRS1 (PS1) is the first 1.8m telescope of the Pan-STARRS project (Kaiser et al. 2010), and the 3π Steradian Survey is one of the primary surveys covering 75% of the sky (Chambers et al. 2016). The PS1 survey provides photometry in five *grizy* bands with limiting magnitudes of 23.3, 23.2, 23.1, 22.3, and 21.4, respectively (Chambers et al. 2016).

We retrieve the photometric data of spectroscopic samples from the public data release 1 (DR1, Flewelling et al. 2020) using the Vizier table[2], which has objects included in the ObjectThin and StackObjectThin tables with nDetections > 2 option. For given positions of spectroscopic samples, we search the nearest object in the photometric table using the search radius of 0.5 arcseconds with the condition of *ObjectQualityFlags* == *QF_OBJ_GOOD* (i.e., good-quality measurement in the PS1) (Flewelling et al. 2020). A single photometric object can be matched to multiple spectroscopic samples. We use the median redshift if more than

[2]`http://cdsarc.unistra.fr/viz-bin/cat/II/349`

two spectroscopic samples correspond to a single PS1 DR1 object. We use the average of two redshifts if two spectroscopic samples are matched to a single photometric object and their redshift difference is less than or equal to 0.005. Objects with a difference of greater than 0.005 are excluded from training samples.

We purposely restrict our training input data to color-related features that allow the easy interpretation of the model results rather than exploring all possible combinations of input features (D'Isanto et al. 2018). The input features comprise four colors $(g - r)$, $(r - i)$, $(i - z)$, and $(z - y)$ in PSF measurement, their uncertainties derived in the quadrature rule, and the same quantities in Kron measurement. The training data include objects only if all four colors are valid in the data release.

Magnitudes of samples are excluded as input; rather, we include the color and color difference uncertainties that implicitly depend on the magnitude and SED of sources. As expected, fainter objects have larger uncertainties. Although uncertainties implicitly contain information correlated with the source magnitudes and SEDs, how NN models treat these inputs may be different.

Furthermore, the $E(B - V)$ value is one of the input features. It is common to apply Galactic dust extinction correction to photometric data as a pre-processing step before photometric redshifts are estimated by machine learning methods. In our case, we let the machine learning models consider the effect of the Galactic dust extinction in the training stage (Beck et al. 2021). We decide to use the $E(B - V)$ values based on the dust emission model of the Plank cosmic microwave background observation (Planck Collaboration et al. 2014), which provides a wide coverage of the sky.

The input features are transformed to improve the performance of the machine learning model. In training machine learning models, data pre-processing is required when input features have different ranges (Sola & Sevilla 1997). Diversely ranged features have a different effect on the loss function adopted in a machine learning model, thus making results biased. From the geometrical standpoint, the data points of differently ranged features form a multidimensional asymmetric volume in the input space. Data pre-processing handles this geometrical asymmetry and makes data more symmetric. Moreover, it smooths out pointy regions that

might exist on the surface of the volume. Our data pre-processing restricts each feature to a comparable range. We test two data pre-processing methods: min-max normalization and standardization. Both methods process data in a feature-wise manner. The former method uses the minimum and maximum values of each feature to restrict feature ranges. For standardization, each feature is re-scaled to have zero mean and unit variance.

After the pre-processing step, we randomly select 80% of our samples as the training set and each 10% of samples as the test and validation sets, respectively. Because we randomly split the data and have many samples, we assume that the samples allocated to each set are drawn from the same distribution. Figure 3.1 shows the redshift distribution of the training, test, and validation sets. To avoid any confusion, we sometimes refer to redshifts as targets in training the machine learning models.

3.3 Method

In this section, we describe the baseline models based on regression, our machine learning approach, ensemble methods for better machine learning performance, and metrics to quantitatively assess the photometric redshift quality.

3.3.1 Baseline Models Based on Regression

For both classification and regression, K-nearest neighbors (KNN) is a relatively simple non-parametric model that can be used (Altman 1992). Because of its simplicity, the model has been used to estimate the photometric redshifts of galaxies and quasars or used as a baseline to evaluate the performances of novel approaches (Zhang et al. 2013; Pasquet-Itam & Pasquet 2018). The model estimates targets (or labels) of given samples based on the averaged target values of k nearest samples in a training set.

Random forest (RF) is an ensemble method using multiple decision trees (Breiman 2001). The model has been extensively used for astronomical classification, regression, and other tasks (Zhang & Zhao 2015). RF is essentially optimized to use feature-based inputs because it recursively splits high-dimensional features by generating multiple root nodes and their succeeding

child nodes. RF, in addition to KNN, is frequently used as a baseline because of its usage of ensemble learning resulting in statistical robustness and its split-rule-learning characteristics, which can return the importance of input features.

A NN is a representative machine learning model inspired by the functioning of the human brain (Hopfield 1982). Generally, NN architecture comprises one input layer, multiple hidden layers, and one output layer. Moreover, these layers are composed of artificial neurons or perceptrons (Rosenblatt 1958), which are the basic units forming the model and mimicking biological neurons. Artificial neurons are interconnected to ones in adjacent layers with randomly initialized connection weights. Vectors provided to neurons in the input layer are transformed by neurons in the hidden layers using connection weights with nonlinear activation functions and are then transmitted to the output layer. Then, neurons in the output layer finally produce scalar or vector outputs. Although the definition of the term is not stringent, the network is usually referred to as a vanilla NN (VNN) when the model generates a scalar output. We emphasize that the regression NN, hereafter, is referred to as VNN because, in our study, it produces a scalar output — a photometric redshift.

We use the mean squared error (MSE) as a loss function to train the parameters of RF and VNN for redshift regression. KNN does not require a loss function. Moreover, we also have tested the adaptive robust loss function proposed by Barron (2019), thus reflecting the anchor loss to be explained later. However, we found that a simpler MSE outperforms the loss in this study.

3.3.2 Multiple-Bin Regression with NN

The MSE for regression problems with the heterogeneous target is not an optimal option (Liu 2019). MSE is the most commonly used loss function to train machine learning models for regression because minimizing the MSE is generally identical for maximizing log-likelihood from a probabilistic standpoint. In most real-life cases, however, the target is heterogeneous rather than homogeneous. If it is the case, using the MSE may result in an undesirable performance of the regression model because the loss function fosters the model to minimize the error throughout all modes while not considering the multi-modal behavior of the target.

Because our target, i.e., redshift, can have a multi-modal distribution particularly because of the degeneracy of redshifts in terms of input features, we consider multiple-bin regression with a NN (hereafter, MBRNN) to bypass the limitation of the MSE for a heterogeneously behaving target. The MBRNN model has been previously explored in several studies for viewpoint estimation (Su et al. 2015; Francisco Massa & Aubry 2016), bounding-box estimation (Mousavian et al. 2017), pose estimation (Kundu et al. 2018), and redshift estimation in astronomy (Pasquet-Itam & Pasquet 2018; Pasquet et al. 2019). Compared with regression using the MSE, these studies demonstrated the performance enhancement of the approach. Furthermore, the property of the probabilistic model enables deeper scrutiny on the causes of poor inference performance for specific samples and examination of model calibration, which are discussed in Section 3.4.1.

We first discretize spectroscopic redshifts and divide them into n independent bins. We test two types of redshift bins: uniform and non-uniform[3]. For uniform bins, we equally discretize the redshifts of the training data with a constant bin width. However, we make each bin have an almost uniform number of samples in the non-uniform binning. Because most objects reside in low redshift ranges, the non-uniform bin width becomes wider as the redshift increases even though it is not monotonic.

The MBRNN model using the softmax function estimates probabilities p_i that the photometric redshifts of objects lie in i^{th} redshift bin. That is, we modify the regression problem into a classification problem with multiple redshift bins. For point estimation, we can compute the photometric redshift z_{phot} either by selecting a redshift bin center with peak probability (i.e., mode) as z_{mode} or by averaging with the output probabilities and central values of the bins as z_{avg} as follows:

$$\begin{gathered} z_{mode} = c_j \text{ for } j = \underset{k}{\operatorname{argmax}}(p_k), \\ \text{or} \\ z_{avg} = \sum p_i \, c_i, \end{gathered} \tag{3.1}$$

where c_i is the central value of i^{th} redshift bin. We compare the prediction accuracy of the two

[3]Some previous studies have considered overlapping bins; however we did not find any advantages of using this strategy for our purpose.

different point-estimation methods in Section 4.3.

We use the anchor loss (Ryou et al. 2019) as a classification loss function for training the model[4]. The loss is designed to measure the disparity of two given probability distributions considering prediction difficulties, which can be attributed to various reasons such as the scarcity of data or the similarities between samples drawn from different distributions. This function evaluates the prediction difficulty using the difference between network-estimated probabilities for the true and other classes. In an easy prediction case, the network-estimated probability of the true class is higher than those of the other classes, whereas it is lower in a difficult case. The difficulty is used to weigh the loss of a sample. For the given two discrete probability distributions, g and p, the anchor loss $\ell(g, p)$ is defined as follows:

$$\ell(g, p) = -\sum_{k} g_k \, log(p_k) + (1 - g_k)(1 + p_k - p_*)^{\gamma} log(1 - p_k), \tag{3.2}$$

where g_k and q_k represent the true and network-estimated probabilities for class k, respectively; p_* represents the anchor probability which means the network-estimated true class probability; γ represents an exponent governing the weights of prediction difficulties. Note that the anchor loss approaches a binary cross-entropy loss, which is one of the most commonly used classification losses, as γ goes to 0.

The baseline VNN and our proposed model MBRNN share the same NN structures except for output layers because these networks generate differently shaped outputs. The networks are composed of fully connected layers only: an input layer with 17 neurons, eight hidden layers sequentially with 128, 256, 512, 1024, 512, 256, 128, and 32 neurons, and the output layer with the number of neurons corresponding to the output size. Each layer is followed by a batch normalization layer (Ioffe & Szegedy 2015) and softplus activation function (Zheng et al. 2015).

3.3.3 Ensemble of multiple-bin method

Our adoption of ensemble methods combines the outputs of multiple MBRNN models and generates an integrated model. Ensemble methods have been extensively used in the field of

[4]Although Ryou et al. (2019) reported that they obtained the highest performance using sigmoid output, we stick to softmax output for MBRNN since we found that the softmax performed better in our case.

machine learning to overcome the limitations of models trained for a limited set of tasks. This strategy often results in better or more generalized model performance by weighting the models depending on their performance or properly assigning each model to the task where the model performs best.

To identify a suitable ensemble approach for our purpose, we evaluate four different methods: plain model averaging ensemble (E1), weighted model ensemble (E2), bin ensemble (E3), and bin-wise selective ensemble (E4) (Yu-yan 2010). E1 is simply averaging results from each model. For E1, the performance of the integrated model might be downgraded when a poorly performing model is included because this method does not consider the specialty of each model. E2 uses weighted averaging of predictions from each model. Assigning higher and lower weights to high- and low-performance models, respectively, this method is frequently expected to yield higher performance than a single model or E1. Optimal weights can be found by various methods such as Bayesian optimization methods (Snoek et al. 2012) and gradient descent methods (Bottou 2010). The gradient descent method with the validation set is used in our study. Moreover, we consider another weighted ensemble method E3, which allocates individual weight to each model and redshift bin combination. Because it provides different weights to each bin in addition to models, this approach has additional parameters to be tuned and more flexibility than E2. Finally, we evalute the selective ensemble method E4, which selects a single model for each redshift bin where the model shows the highest point estimation accuracy using the validation set. In E4, we find redshift bins to which point-estimated redshifts of each test object belong using the vote of the single models. Then, the model allocated to the bin is used to estimate probability distributions for objects.

3.3.4 Metric

We use multiple metrics to assess performance in estimating photometric redshifts. We refer to Cavuoti et al. (2017) for the detailed description of metrics. The following is a brief explanation of metrics:

- *Bias*: the absolute mean of redshift differences defined by $\frac{1}{N}|\sum_i^N \Delta z_i|$ where N represents the number of samples in the dataset, i represents the sample index, and $\Delta z_i = (z_{i,spec} -$

Table 3.2 Metrics comparison between the average and mode point estimation of the MBRNN and baseline models. The best performing cases in each search are presented in boldface.

Model	Bias	MAD	σ	σ_{68}	*NMAD*	R_{cat}
MBRNN (mode)	0.0036	0.0266	0.0430	0.0274	**0.0256**	0.0111
MBRNN (avg)	**0.0017**	**0.0254**	**0.0394**	**0.0272**	0.0256	**0.0084**
KNN	0.0055	0.0324	0.0487	0.0352	0.0340	0.0158
RF	0.0027	0.0277	0.0430	0.0294	0.0277	0.0116
VNN	0.0046	0.0307	0.0450	0.0338	0.0327	0.0113

$z_{i,phot})/(1 + z_{i,spec})$,

- *MAD*: the mean of absolute differences defined by $\frac{1}{N}\sum_i^N |\Delta z_i|$,
- σ: the standard deviation of the difference $z_{i,spec} - z_{i,phot}$,
- σ_{68}: the 68th percentile of the absolute difference, i.e., $|z_{i,spec} - z_{i,phot}|$,
- *NMAD*: the normalized median absolute deviation of the differences, which is $1.4826 \times Median(\Delta z)$,
- R_{cat}: catastrophic (hereafter, *cat*) error, which corresponds to $|\Delta z| > 0.15$, fraction.

Using these metrics, we evaluate the quality of photometric redshifts and perform a grid search to identify the empirically optimal configuration of the MBRNN model. For the numerical metrics, the lower the values, the better the model performance.

3.4 Result

3.4.1 Single Model Performance Test

We quantitatively assess and visually inspect photometric redshifts obtained by the MBRNN model using test set samples. The empirically optimal model configuration for inference is selected in the grid search (see appendix 3.A). The default model setup adopts the anchor loss

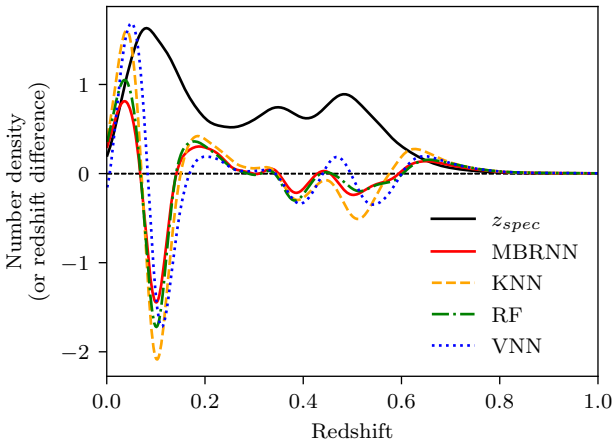


Figure 3.2 Redshift difference distributions of the MBRNN and baseline models. The distribution of spectroscopic redshifts is also drawn here for comparison. The spectroscopic redshift distribution is scaled down by half for visual clarity. Positive and negative differences indicate under- and over-estimation of the distribution, respectively.

with γ of 0 and 64 uniform redshift bins. Henceforth, unless otherwise stated, we shall restrict ourselves to the MBRNN model with this setup.

Because the baseline models perform regression, we require a way to derive the point estimation of redshifts with the MBRNN model for a fair comparison with the baseline models. We first juxtapose the metrics of mode and average photometric redshifts, as shown in Table 3.2, to determine which estimation is more suitable for point estimation. Average redshifts evince lower metrics than mode redshifts. Because this result indicates that the average estimation accompanies higher point estimation accuracy than the mode estimation, we use average redshifts in the rest of the analysis related to point estimation, except cases specifying the usage of the mode redshift. Using the average estimation of redshifts, we now can equitably compare the MBRNN model with the baseline models.

The MBRNN model shows a higher prediction accuracy for overall redshift ranges than baseline models. Figure 3.2 shows the distribution of differences between spectroscopic and photometric redshifts in the MBRNN and baseline models. We obtain the distributions of the differences by subtracting the distribution of photometric redshifts from that of spectroscopic

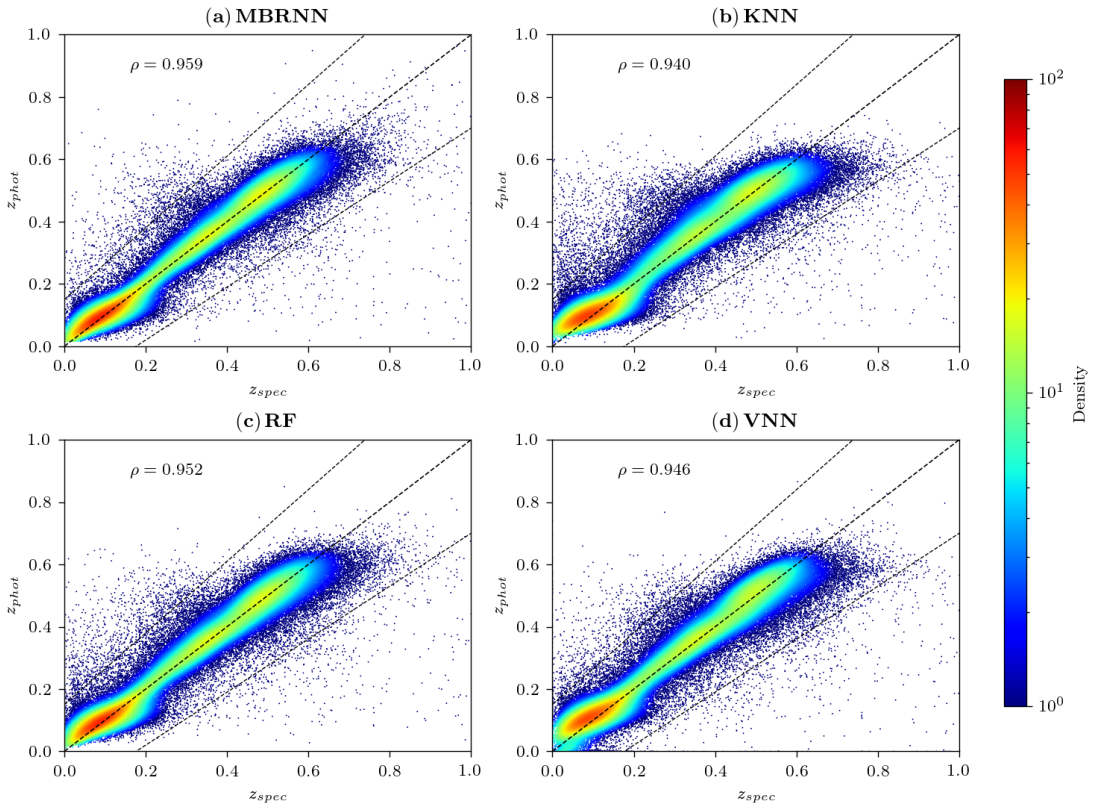


Figure 3.3 Distributions of spectroscopic and photometric redshifts obtained by the MBRNN and baseline models. The markers are color-coded with number density, and the color is normalized in the log scale for better contrast. Samples outside the dashed lines on both sides correspond to *cat* objects, and the central dashed line has a slope of 1. The Pearson correlation coefficients ρ between spectroscopic and photometric redshifts are also presented in each panel.

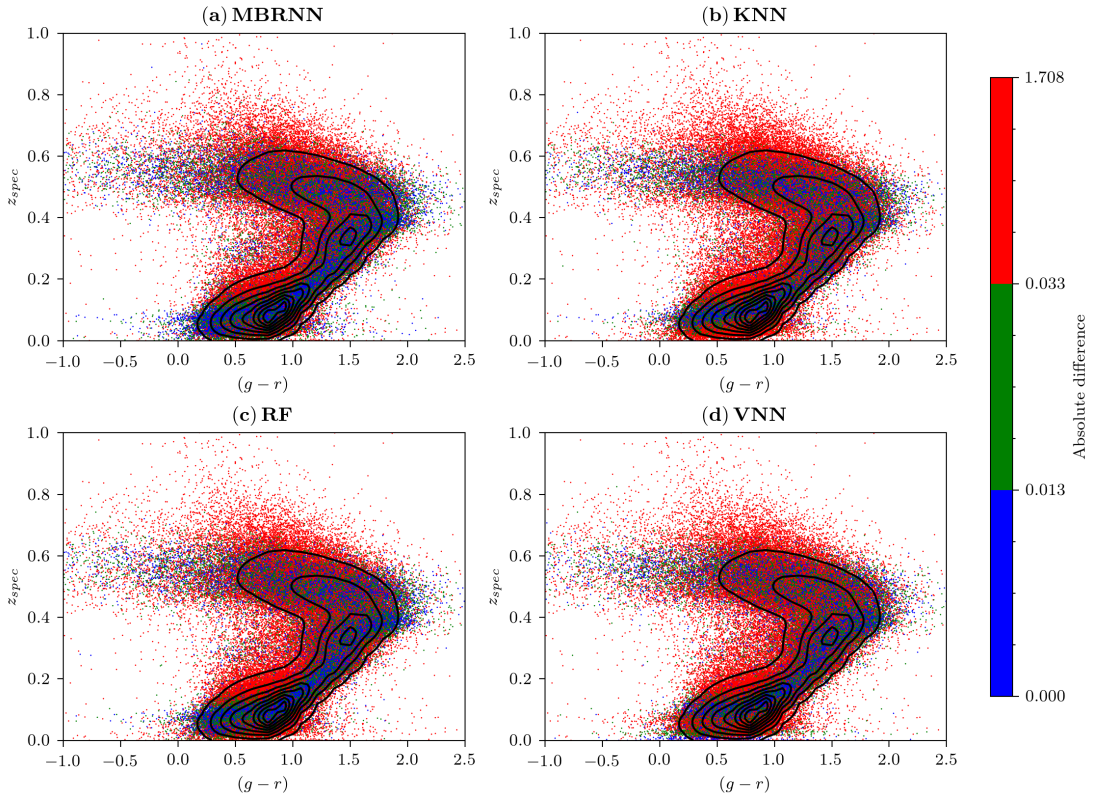


Figure 3.4 Distribution of the absolute difference between spectroscopic and photometric redshifts in the space of $(g - r)$ in Kron measurement and spectroscopic redshifts. The contour lines show the densities of the samples in this space. We assign samples to low-, middle-, and high-difference groups, and the samples in each group are represented in different colors. Providing the smallest area of the high-difference region in the MBRNN model compared with the other baseline models, the MBRNN model reproduces spectroscopic redshifts with the highest accuracy in the outskirt area of the contours.

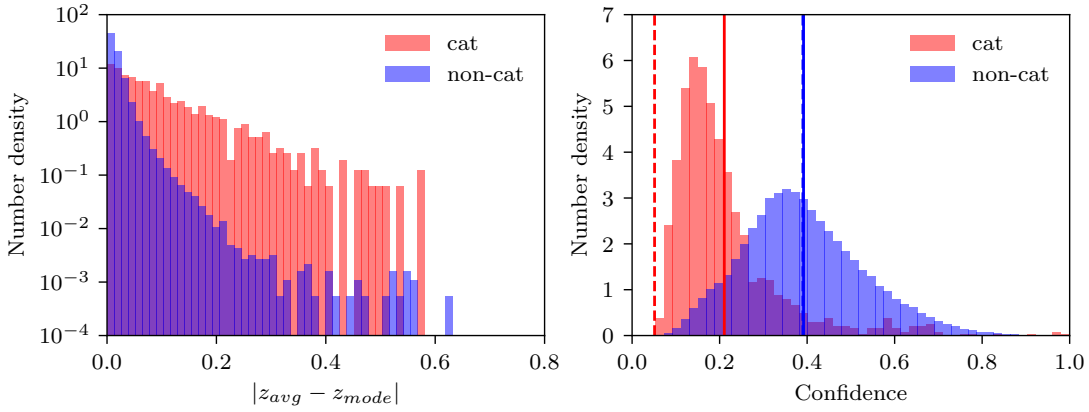


Figure 3.5 Distribution of the difference between the average and mode photometric redshifts for the *cat* and *non-cat* cases. *Right*: Distribution of the confidence estimation for photometric redshifts. The vertical solid and dashed lines mark the mean confidence and accuracy, respectively, for each case.

redshifts. The MBRNN model's relatively close-to-zero differences indicate that it best captures the multi-modal behavior of targets, i.e., redshifts. The KNN, RF, and VNN models show poorer descriptions for the heterogeneous property of targets, particularly for redshift ranges with distribution peaks. These results are naively expected from the lowest metrics of the MBRNN model compared with the baseline models as presented in Table 3.2, even though the metrics, which are marginalized into one scalar throughout multiple dimensions, cannot represent locally different behaviors of the distribution.

The distributions of spectroscopic and photometric redshifts shown in Figure 3.3 summarize well the differences reported in the metrics. The density peaks of the MBRNN and RF models are aligned with the slope-one line, whereas those of the KNN and VNN models are misaligned. This difference reflects lower deflection-related metrics (i.e., bias and MAD) of the MBRNN and RF models. As expected from the dispersion-related metrics (i.e., σ and σ_{68}), the redshift distribution of the MBRNN model shows the smallest dispersion around the slope-one line. This characteristic is conspicuous in redshift regions with the distribution peaks (i.e., $z_{spec} \sim 0.125$, 0.35, and 0.5), as presented in Figure 3.1, where scatters spread more extensively than other redshift regions. The small bias and dispersion values of the MBRNN model

contribute to the highest Pearson correlation coefficient.

The lack of objects in the input space is one of the primary causes that induce prediction difficulty. Figure 3.4 shows the distribution for the absolute difference of redshifts between spectroscopic and photometric redshifts in the space of the input $(g - r)$ in Kron measurement and target redshifts. Although this visualization only shows the projected distribution in the input $(g - r)$ space with degeneracy of the other input features in the higher dimensional space, the $(g - r)$ color has a well known correlation with redshifts, and its interpretation is usually straightforward in examining models (e.g. Korytov et al. 2019). When grouping the samples into low-, middle-, and high-difference groups corresponding to about 33% of the samples per group in the MBRNN model, a large number of samples with significant redshift discrepancy are found independently in models. Intriguingly, the high-difference groups are similarly distributed in every model, whereas the area of the regions corresponding to this group is the smallest in the MBRNN model. These regions are the outskirts of the density contour lines in which the objects sparsely reside. As presented in Appendix 3.B, the distribution of the *cat* samples in the MBRNN model is similar to that of baseline models. These model-independent patterns also can be found in Figure 3.2. The results indicate that the samples populating over the specific regions in the input space are likely to bring high prediction difficulty regardless of the model.

The high-difference samples possibly have multi-modal model probability distributions. Benefiting from the MBRNN model's property as a probabilistic classification model, we examine the distribution of the difference between the average and mode photometric redshifts for the *cat* and *non-cat* samples. Figure 3.5 shows that the *cat* error samples have higher differences than *non-cat* ones. The higher differences between the average and mode estimation of the *cat* samples indicate that the model-estimated probabilities of *cat* samples are possibly multi-modal.

Furthermore, the confidence distribution of the MBRNN model shown in Figure 3.5 endorses our interpretation of the *cat* samples. Confidence expresses a model's level of certainty about the classification result for a given sample. The measure of confidence is defined as the maximum value in the probability output of a model (Hendrycks & Gimpel 2017). The low

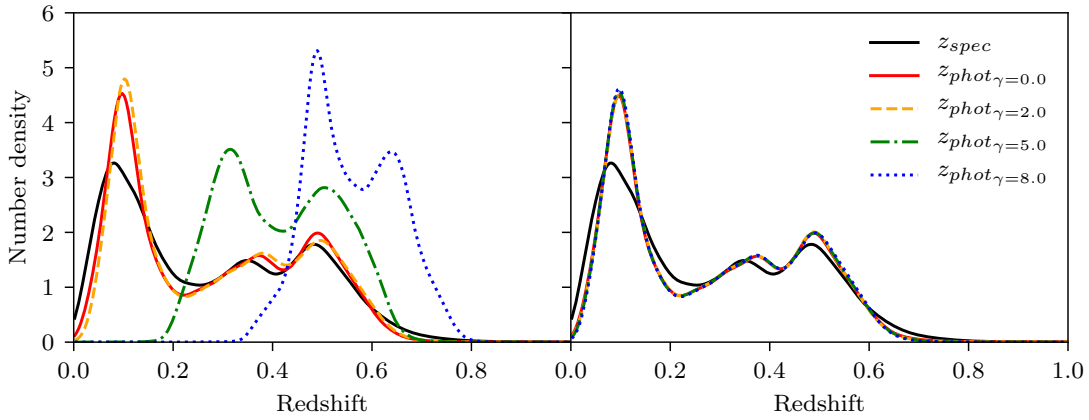


Figure 3.6 Distribution of photometric redshifts in the MBRNN model with 64 uniform bins (*left*) and 512 non-uniform bins (*right*) with the anchor loss γ of 0, 2, 5, and 8. The distribution of spectroscopic redshifts is also drawn for comparison.

confidence of *cat* samples indicates a high possibility that these samples have multiple probability peaks.

It is worthwhile to refer to the practical guide for flagging objects that require caution in analysis. In the confidence histogram presented in Figure 3.5, the difference between mean confidence and accuracy of *cat* samples is higher than that of *non-cat* samples. We may refer that the model is well-calibrated for *non-cat* samples because mean confidence and accuracy are comparable (Guo et al. 2017). However, the model is overconfident about the *cat* samples because the mean confidence of the model is significantly higher than mean accuracy. In this case, caution is required because overconfident estimation leads to high prediction error and may lead to an incorrect interpretation. We recommend being cautious about objects outside the dense regions in the input space because the model is overconfident about *cat* samples, and most of the *cat* samples reside in the low-density region of input space (see Figure 3.4).

3.4.2 Ensemble Model Performance Test

Performance elevation of the integrated model through ensemble learning emerges from diverse specialties of individual models. During grid-search, we find that the anchor loss assigns different specialties to models based on γ. As the value of γ increases, the distributions of pho-

Table 3.3 The metrics of the ensemble learning. We also summarize the results of the single model for comparison.

Bin type	Case	Bias	MAD	σ	σ_{68}	*NMAD*	R_{cat}
64 uniform bins	Single Model	0.0017	0.0254	0.0394	0.0272	0.0256	0.0084
	E1	0.0023	0.0255	0.0393	0.0274	0.0258	0.0084
	E2	0.0019	0.0254	0.0392	0.0273	0.0256	0.0083
	E3	**0.0010**	**0.0253**	**0.0389**	**0.0272**	**0.0255**	**0.0082**
	E4	0.0020	0.0255	0.0394	0.0273	0.0257	0.0086
512 non-uniform bins	Single Model	0.0023	0.0256	0.0404	0.0272	0.0255	0.0090
	E1	0.0021	0.0255	0.0400	0.0272	0.0255	0.0088
	E2	0.0021	0.0255	0.0400	**0.0272**	0.0255	0.0088
	E3	**0.0021**	**0.0255**	**0.0400**	0.0272	**0.0255**	**0.0087**
	E4	0.0022	0.0256	0.0405	0.0272	0.0255	0.0090

tometric redshifts shift toward the higher redshift region in the uniform bin case (see Figure 3.6). Because the loss with larger γ allocates more difficult objects for prediction with bigger weights, this model bias is attributed to the higher redshift region which corresponds to the most difficult part for prediction from the model's perspective [5].

We examine four ensemble methods explained in Section 4.2. Because the models with high anchor loss γ focus excessively on the high redshift region, we perform ensemble learning with models trained with moderate values of γ: 0, 0.2, 0.5, and 1. Furthermore, for comparison, we test ensemble methods with 512 non-uniform bins trained with the same set of γ values. A set of models used for each ensemble combination has the same architecture and number of redshift bins. The ensemble experiments with different numbers of bins and sets of γ values can be found in Appendix 3.C.

[5]It is anticipated because our dataset rarely has high-redshift samples, and hence the equal-width bins in the high-redshift region have a comparably small number of objects. Besides, we have already confirmed that data sparsity contributes to the prediction difficulties in Section 3.4.1.

Table 3.4 Comparison data with either spectroscopic or photometric redshifts.

Name	Number of objects	Maximum redshift	Reference
HeCS	20,544	0.72	Rines et al. (2013)
SDSS-DR12-LPSC	38,818	1.00	`https://www.sdss.org/dr12/algorithms/photo-z/`
HSC-PDR2-Mizuki-Galaxy	6,996	1.47	Nishizawa et al. (2020)
HSC-PDR2-Mizuki-NonGalaxy	3,267	4.15	

We report that E3 with 64 uniform bins has the best performance metrics. Moreover, the E3 ensemble model is well-calibrated. Appendix 3.D provides the calibration study of the ensemble model. In other words, E3 successfully considers the varied specialties of the individual models with different values of γ. Table 3.3 compares the results of the four different ensemble methods using 64 uniform bins and 512 non-uniform bins.

For the 512 non-uniform bin cases, all methods outperform the single model; however, the performance disparity between the ensemble methods is not as pronounced as it is in the uniform bin case. We recognize that it is because improvement is attributed to the stochastic differences between models, and not diverse specialties. The non-uniform bins are designed to make each bin have an almost uniform number of samples. Since the prediction difficulty mostly stems from the scarcity of data, the non-uniform bin cases have no particularly challenging parts to predict. Therefore, it results in monotonous specialties of individual models being insensitive to the variation of γ. Figure 3.6 shows that the redshift distributions in the MBRNN model with the non-uniform bins do not vary significantly as the γ increases. Consequently, it results in a small improvement in all ensemble methods because of the stochastic difference among the single models.

3.5 Model Validation

We compare our photometric redshift results with other existing spectroscopic or photometric redshifts from various sources to investigate when the trained models fail or are not trustworthy. In particular, the data collected from the OOD with respect to the training data distribution should have highly uncertain photometric redshifts because the training data have never allowed the model to acquire the relevant information about the OOD data (e.g., Ren et al. 2019).

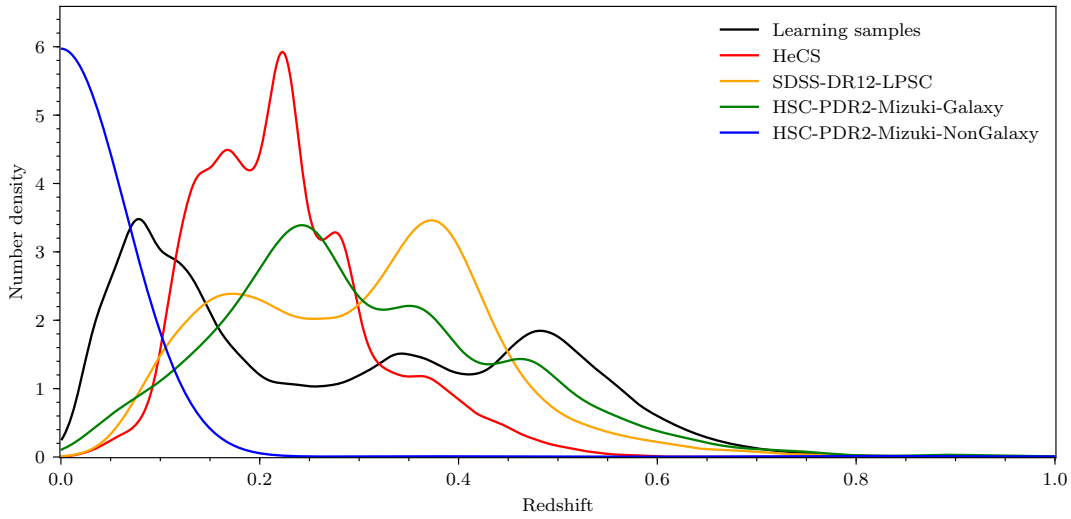


Figure 3.7 Redshift distribution of the comparison samples for model validation as summarized in Table 3.4. The plot displays redshifts up to approximately 1 even though the maximum redshift is 4.15 in the HSC-PDR2-Mizuki-NonGalaxy.

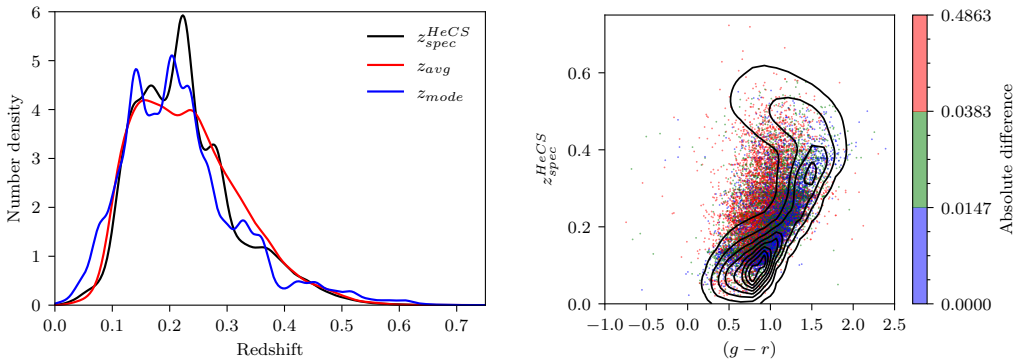


Figure 3.8 *Left*: Distribution of the comparison HeCS spectroscopic redshifts and derived photometric redshifts z_{avg} as the average value, and z_{mode} as the mode value. *Right*: The distribution of the absolute difference between the HeCS spectroscopic redshifts and the derived z_{avg} in the space of $(g-r)$ and HeCS spectroscopic redshifts. Following Figure 3.4, the samples are color-coded, and the distribution of the training samples is drawn as contour lines for comparison. The distribution of the absolute difference between z_{spec} and z_{avg} with respect to $(g-r)$ in Kron measurement evidently shows that the incorrect estimation of z_{avg} is due to the lack of sufficient training samples to occupy the input space.

Table 3.4 summarizes the comparison data, and Figure 3.7 shows the redshift distribution of the comparison samples. We match the PS1 DR1 data to comparison data with a search radius of $0.5''$, and we apply the same selection and filtering rules to the PS1 DR1 data as we do to the training samples.

3.5.1 Validation with Spectroscopic Redshift Samples

We estimate the photometric redshifts of the matched PS1 DR1 objects for the dataset HeCS. The HeCS dataset includes objects with the spectroscopic redshifts of galaxies found in galaxy cluster areas over z = 0.1 – 0.3 (Rines et al. 2013). Therefore, this dataset contains both cluster members and line-of-sight field galaxies. We use objects with a redshift quality flag of Q, thus indicating a secure redshift. Moreover, we exclude objects with redshifts < 0.009 because they are not galaxies generally.

The comparison with spectroscopic samples allows us to unbiasedly evaluate the performance of our trained machine learning model although the comparison sample size is not as large as that of training samples. As shown in Figure 3.7, the true redshift range is well matched to the redshift range over which we intend to use the trained model. However, the distribution of true redshifts differs from that of training samples. Moreover, this comparison helps us assess how well the trained model can be used in idenitfying potential galaxy groups and clusters with the PS1 DR1 data (see Euclid Collaboration et al. 2019, for discussion).

As shown in Figure 3.8, the photometric redshifts are well estimated for the redshift range of the HeCS dataset. However, we find that certain objects with the biased estimation of photometric redshifts over $0.1 < z < 0.3$. As the distribution of the absolute difference between the spectroscopic and photometric redshifts in the space of $(g - r)$ and the spectroscopic redshifts highlights, the biased estimation of the photometric redshifts is attributed to the lack of blue training samples for given redshifts over $0.1 < z < 0.3$. Although the distribution in the input space such as $(g - r)$ does not completely show the distribution mismatch with respect to the model because of the model's nonlinearity, the mismatched training and test HeCS distributions lead to the model's biased results. This result indicates that our model may require to be updated with sufficient samples of blue galaxies to estimate photometric redshifts in galaxy

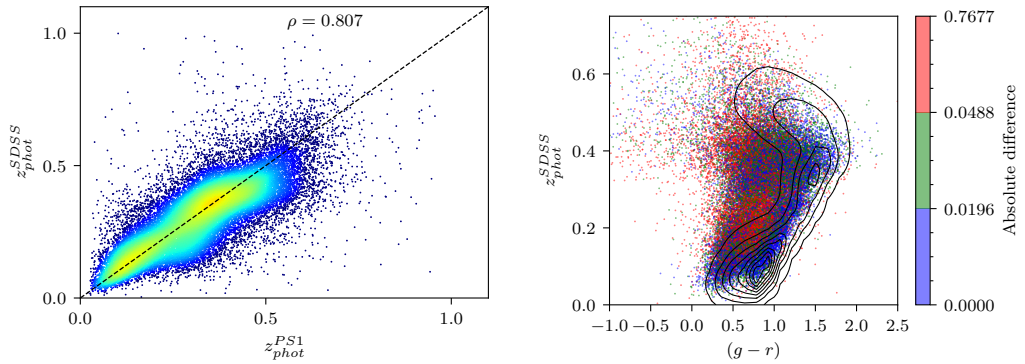


Figure 3.9 *Left*: Distribution of the comparison SDSS photometric redshifts and derived photometric redshifts z_{avg} for the PS1 photometric data. *Right*: The distribution of the absolute difference between SDSS photometric redshifts and the derived z_{avg} in the space of $(g-r)$ (Kron measurement) and SDSS photometric redshifts. The scatters are color-coded, as in Figure 3.8, and the distribution of PS1 training samples are drawn as contour lines for comparison. The distribution of the absolute difference between SDSS photometric redshifts and the derived z_{avg} with respect to $(g-r)$ shows that the significant fraction of samples with a large difference of photometric redshifts have $(g-r)$ out of the main training samples. However, we still find some consistent estimation of photometric redshifts between the two different methods/data for $(g-r) \sim 0.7$ and $z_{phot}^{SDSS} \sim 0.3$.

groups and clusters over $0.1 < z < 0.3$.

3.5.2 Validation with Photometric Redshift Samples

Our trained model is compared with other photometric redshift estimation methods. Because there are more galaxy objects with known photometric redshifts than those with spectroscopic redshifts, we intend to use this comparison to assess the validity of the trained model over a wide range of redshifts and input space although photometric redshifts are not as precise and accurate as spectroscopic redshifts.

As summarized in Table 3.4, we use the SDSS Data Release 12 (DR12) (Alam et al. 2015), which contains photometric redshifts derived from a nearest-neighbor fit with a kd-tree structure of training samples (Csabai et al. 2007). We extract SDSS photometric redshifts for objects

found in the area centered at RA 26.25°, DEC -7.5°with a radius of 2.5°, and we extract the PS1 DR1 objects corresponding to the SDSS objects that were identified. After discarding objects with SDSS spectroscopic redshifts, the sample size becomes 44,890 for the same filtering rule that we use for the training data of the PS1 DR1 data. We examine point-source scores of the 44,890 objects in the PS1 DR1 catalog (Tachibana & Miller 2018), leaving out only extended sources (i.e., galaxies) in the PS1 image data with the condition of the point-source score ¡ 0.1. The filtered data, including 38,818 objects commonly found in the SDSS and PS1, should comprise only galaxy-like objects for which our trained model should be able to estimate photometric redshifts.

Generally, two different estimations of photometric redshifts are consistent although the methods and training data are completely different. Figure 3.9 shows the comparison between the two photometric redshift inferences where z_{phot}^{PS1} corresponds to z_{avg} in our estimation and z_{phot}^{SDSS} represents estimation by robust fit to nearest neighbors in the SDSS reference set. Pearson's correlation coefficient $\rho \sim 0.81$ confirms that the two methods are consistent for most samples. A dominant fraction of objects with discrepant redshifts seems to have a range of colors not properly represented by training samples; however, we report certain fraction of objects to have consistent redshifts although their colors do not follow the color of the training samples (see Figure 3.9).

We examine possible causes of certain objects showing a large discrepancy in the estimated redshifts. For example, `SDSS J014232.69-090324.5` corresponding to PS1 DR1 objid = 97130256361222090 seems blended in the PS1 DR1 and the SDSS images. The photometric redshifts of this object are 0.203 and 0.858 as z_{phot}^{SDSS} and z_{phot}^{PS1}, respectively. The `SDSS J014228.47-072343.0` (i.e., PS1 DR1 objid = 99120256186386095) shows possible effects of blending in their images with the photometric redshifts $z_{phot}^{SDSS} = 0.018$ and $z_{phot}^{PS1} = 0.391$. The blending of sources is a well-known problem for estimating photometric redshifts (D. M. Jones & Heavens 2019; LSST Dark Energy Science Collaboration (LSST DESC) et al. 2021). We also find certain objects have conflicting colors between the SDSS DR12 and PS1 DR1 catalogs. For example, `SDSS J014126.01-065056.2` matched to the PS1 DR1 objid = 99780253584271692 has $(g - r) = 1.1 \pm 0.01$ (Kron color) and 0.7 ± 0.32

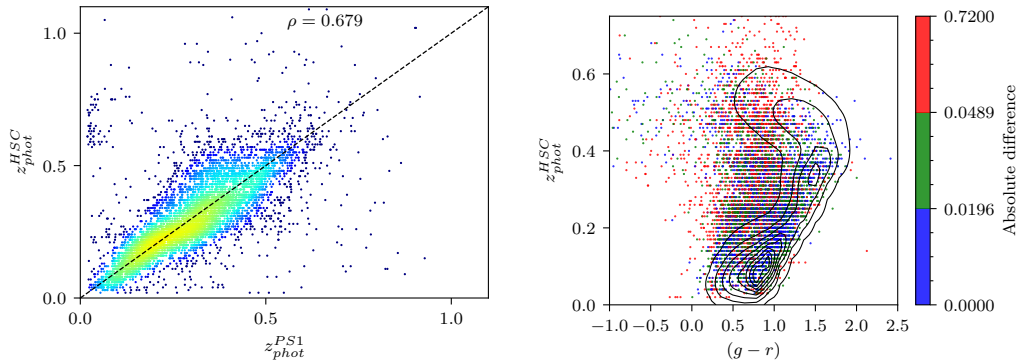


Figure 3.10 *Left*: Distribution of the comparison HSC galaxy photometric redshifts and derived photometric redshifts z_{avg} for the PS1 photometric data. *Right*: The distribution of the absolute difference between the HSC photometric redshifts and the derived z_{avg} in the space of $(g - r)$ (Kron measurement) and HSC photometric redshifts. The color-coding of scatters and contour lines are as those of Figure 3.9. The pattern of the discrete z_{phot}^{HSC} distribution appears clearly in both plots. The correlation ρ between the two photometric redshifts is lower than that in for the comparison with the SDSS photometric redshift presented in Figure 3.9 partly due to the discrete distribution of z_{phot}^{HSC}.

(aperture color) in the SDSS DR12 and PS1 DR1 catalogs, respectively. The photometric redshifts of this object are $z_{phot}^{SDSS} = 0.941$ and $z_{phot}^{PS1} = 0.172$.

3.5.3 Model Outcomes for Non-galaxy Objects

The validity of trained machine learning models depends on the assumption that the training and test data follow the same distribution from the learning model's view. Therefore, when the trained model infers photometric redshifts for objects obtained from the OOD data, the estimation should be highly *uncertain* and/or *inaccurate*. We already present the case showing this effect in Section 3.5.1 for the slightly different distribution of input features for galaxies between training samples and HeCS test data.

The application of the trained model on the datasets HSC-PDR2-Mizuki-Galaxy and HSC-PDR2-Mizuki-NonGalaxy enables us to evaluate the results for the physically OOD objects, i.e., non-galaxy objects. Photometric redshifts in these datasets are results acquired in running

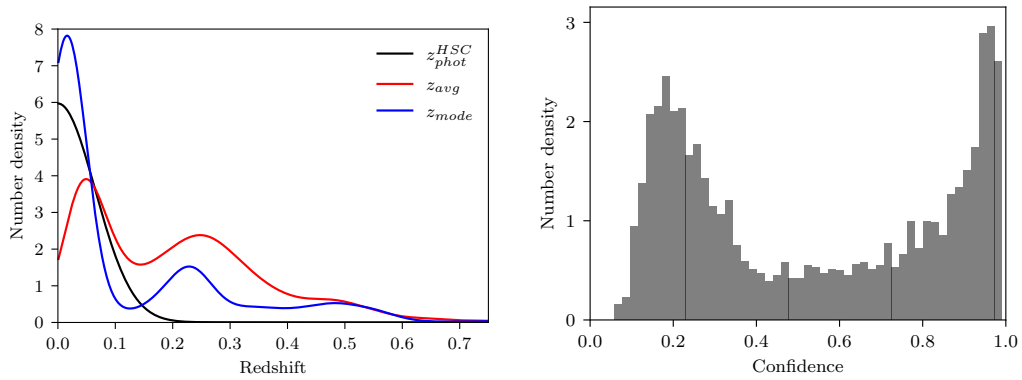


Figure 3.11 Distribution of the comparison HSC non-galaxy spectroscopic redshifts, derived photometric redshifts z_{avg} as the average value, and z_{mode} as the mode value (*left*). Confidence distribution of the non-galaxy samples (*right*) shows that our trained model is overconfident on some of the physically OOD samples.

a template fitting-code MIZUKI (Tanaka 2015; Tanaka et al. 2018; Nishizawa et al. 2020). The estimation products have probabilities of being stars, quasars, and galaxies. We select objects with photometric redshifts around RA 31.25°, DEC -2.5°with a radius of 2.5°. Following the filtering criteria used for the training data, we extract 6,996 objects as galaxy objects and 3,267 objects as non-galaxy objects with the PS1 DR1 data. The number of stars is 3,171 among 3,267 non-galaxy objects. Therefore, ∼100 objects are classified as quasars in the HSC test data.

The estimated photometric redshifts of galaxy objects in the HSC comparison data are closely matched to those esimated by the trained model as shown in Figure 3.10. In general, our machine-learning estimation of photometric redshifts seems consistent with those derived by the template fitting-code. However, we report certain systematic difference patterns such as objects with $z_{phot}^{HSC} \sim 0.6$ for $z_{phot}^{PS1} \sim 0.1$. The discrete distribution of z_{phot}^{HSC} appears to be a systematic pattern embedded in the HSC Mizuki inference of photometric redshifts. As shown in Figures 3.8 and 3.9, we report that most objects with a large difference between the two photometric redshifts can be considered OOD samples based on the input data properties (see Figure 3.10).

We examine how our trained model estimates photometric redshifts of non-galaxy objects.

If the model is well-trained to inductively infer the galactic photometric redshifts, photometric redshifts of the physically OOD samples (i.e., stars and galaxies) should follow the overall redshift estimation of galaxies when the OOD samples have similar input values as training galaxies. The useful machine learning model should result in a highly uncertain estimation of photometric redshifts for data samples from the OOD with respect to the input space and the trained model even when the OOD samples are galaxies.

Figure 3.11 shows the distribution of photometric redshifts for the non-galaxy HSC objects. Because the majority of non-galaxy HSC objects are stars, their redshift distribution has a peak at z = 0. Comparing the distribution to that of the training samples (see Figure 3.7), the derived photometric redshifts of the physically OOD samples share a similar distribution to that of the training samples at z $\sim$ 0.1 and 0.5. However, the derived photometric redshifts of the physically OOD samples show a concentrated distribution around z = 0.25 where the input values and photometric redshifts of these physically OOD samples are distributed in a manner similar to the galaxy OOD samples as shown in Figure 3.10 (i.e., the galaxy OOD objects at $(g - r) \sim 0.7$ and $z_{phot}^{HSC} \sim 0.25$).

As shown in Figure 3.11, the trained model produces overconfident results on certain physically OOD samples. We anticipate that the well-trained model will output nearly identical probability distributions for OOD samples as random guesses, i.e., uniform distribution. In such a case, the confidence distribution should be nearly uni-modal with a peak in the low confidence range and then monotonically diminish as the confidence increases. However, the confidence distribution of physically OOD samples is multi-modal and has the tallest peak with high confidence. Therefore, the high confidence of the trained model's outputs does not guarantee that the tested sample has the same distribution as the training samples, particularly for the physically OOD samples. Unless the physically OOD samples such as stars and quasars are separately classified (e.g., Fotopoulou & Paltani 2018), the application of our model to these OOD samples can produce incorrect inference results with overconfidence.

Table 3.A.1 Metrics of the grid search for the best model configuration of the MBRNN model. As mentioned in Section 4.3, the lower metrics indicate a higher accuracy of the model.

(a) Metrics for different data scaling methods, and the existence of color uncertainties and $E(B-V)$ as input features.

Data scaling	Color uncertainties	$E(B-V)$	Bias	MAD	σ	σ_{68}	*NMAD*	R_{cat}
Min-max	Yes	Yes	**0.0019**	**0.0256**	**0.0398**	**0.0273**	**0.0256**	**0.0088**
		No	0.0021	0.0267	0.0412	0.0287	0.0270	0.0096
	No	Yes	0.0028	0.0316	0.0496	0.0329	0.0310	0.0188
		No	0.0030	0.0336	0.0533	0.0346	0.0327	0.0226
Standardization	Yes	Yes	0.0022	0.0256	0.0399	0.0273	0.0257	0.0086
		No	0.0023	0.0268	0.0415	0.0288	0.0270	0.0095
	No	Yes	0.0030	0.0317	0.0498	0.0329	0.0310	0.0191
		No	0.0031	0.0336	0.0535	0.0346	0.0328	0.0229

(b) Metrics for different anchor loss γ values.

Loss configuration		Bias	MAD	σ	σ_{68}	*NMAD*	R_{cat}
Binary Cross Entropy ($\gamma = 0$)		**0.0019**	**0.0256**	0.0398	**0.0273**	**0.0256**	0.0088
Anchor Loss	$\gamma = 0.2$	0.0025	0.0257	0.0398	0.0275	0.0259	0.0087
	$\gamma = 0.5$	0.0027	0.0256	**0.0396**	0.0274	0.0258	**0.0086**
	$\gamma = 1$	0.0027	0.0256	0.0396	0.0275	0.0260	0.0086
	$\gamma = 2$	0.0039	0.0260	0.0399	0.0280	0.0266	0.0087
	$\gamma = 5$	0.0327	0.0483	0.0611	0.0536	0.0507	0.0451

(c) Metrics for different redshift bin configurations.

Bin type		Bias	MAD	σ	σ_{68}	*NMAD*	R_{cat}
Uniform Bins	32	0.0020	0.0257	0.0399	0.0274	0.026	0.0086
	64	**0.0017**	**0.0254**	**0.0394**	**0.0272**	0.0256	**0.0084**
	128	0.0019	0.0256	0.0398	0.0273	0.0256	0.0088
	256	0.0024	0.0254	0.0396	0.0272	**0.0255**	0.0084
	512	0.0024	0.0255	0.0396	0.0273	0.0256	0.0086
Non-uniform Bins	32	0.0144	0.0337	0.0577	0.0310	0.0281	0.0342
	64	0.0075	0.0285	0.0470	0.0288	0.0266	0.0168
	128	0.0045	0.0266	0.0429	0.0278	0.0259	0.0117
	256	0.0028	0.0259	0.0415	0.0274	0.0255	0.0096
	512	0.0023	0.0256	0.0404	0.0272	0.0255	0.0090

3.A Search for the Optimal Configuration of the MBRNN Model

We perform a grid search by varying model configurations with the validation data to find the empirically optimal configuration of the MBRNN model. The grid search includes three different sets of configuration variations. First, when setting the number of redshift bins to 128 and using the anchor loss with $\gamma = 0$, we examine the changes in the model's accuracy in terms of input data scaling methods (i.e., standardization vs. min-max normalization), and the existence of color uncertainties and $E(B-V)$ as input features.

The model using the min-max normalization, color uncertainties, and $E(B-V)$ shows the best point-estimation accuracy as summarized in Table 3.A.1 (a). Including the color uncertainties as input features has the largest impact on point-estimation accuracy among three factors. The min-max normalization and inclusion of $E(B-V)$ in the input have small positive effects on point-estimation accuracy.

Second, we perform the grid-search for the anchor loss parameter γ fixing the number of redshift bins to 128. Because the anchor loss requires a prior setup of γ before training, we examine a set of γ values of 0 (i.e., binary cross entropy loss), 0.2, 0.5, 1, 2, and 5. The results of this second grid search are presented in Table 3.A.1 (b). The model with γ of 0 outperforms the others in terms of the overall accuracy as naively expected.

Using the maximum performance configuration, we finally examine the various strategies of redshift binning and the number of bins. The search includes a comparison of results between uniform and non-uniform binning methods as well as the results for the 32, 64, 128, 256, and 512 redshift bins. In the non-uniform redshift binning, we set the bin edges such that each bin contains nearly the same number of samples, as explained in Section 3.3.2. As illustrated in Table 3.A.1 (c), the MBRNN model with 64 uniform bins outperforms the other configurations for most metrics, although the difference is not significant. Moreover, we discover that the uniform binning case outperforms the non-uniform one.

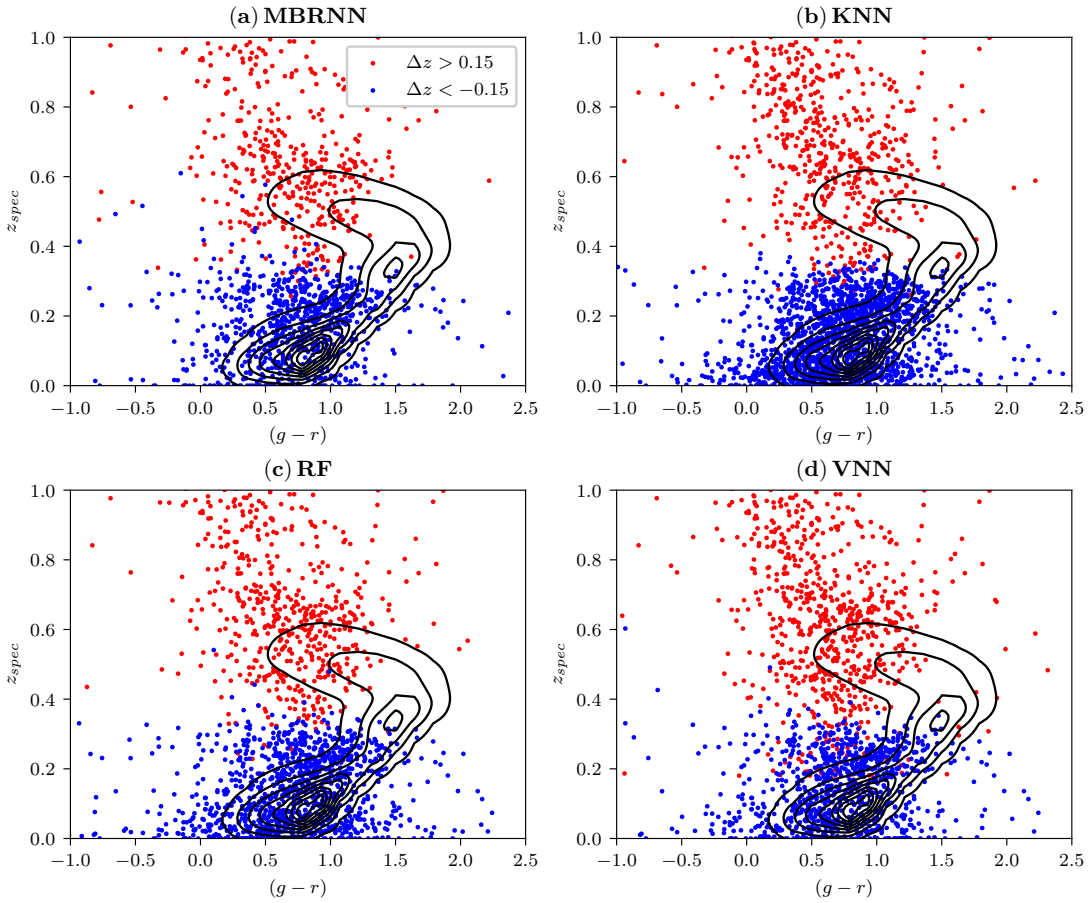


Figure 3.B.1 Distributions of the *cat* samples in the space of $(g - r)$ in Kron measurement and spectroscopic redshifts. The *cat* samples are shown with different colors for $\Delta z > 0.15$ (red) and $\Delta z < -0.15$ (blue). The lower fraction of the *cat* samples in the MBRNN than in other models is mainly due to the reduction of these samples at low redshifts.

3.B Catastrophic Samples

We examine the distributions of the *cat* samples in the input space according to models and find that these samples populate in the similar regions of the input space. Figure 3.B.1 presents the distributions of under and overestimated *cat* samples in the MBRNN and baseline models. As shown in the figure, the point-estimated photometric redshifts of the *cat* samples in MBRNN with true spectroscopic redshifts larger and lower than $z_{spec} \sim 0.4$ tend to be under- and over-estimated, respectively. This pattern appears in the other baseline models too, and the under- and over-estimated photometric redshifts of the *cat* samples are similarly distributed in the input space. However, the area taken by the *cat* samples in the MBRNN model is smallest among the models, being consistent with the fact that the R_{cat} is smallest in the MBRNN model (see Table 3.2).

We also find that MBRNN works better than the baseline models particularly in the low redshift range. Specifically, we compare how many *cat* samples found in the RF model, which shows the best performance among the baseline models, become *non-cat* samples in the MBRNN model. The RF model has 766 *cat* samples in its test, and the entire 543 and 223 samples with over- and under-estimated photometric redshifts in the RF model are the *non-cat* samples in the MBRNN. Especially, samples with $z_{spec} \sim 0$ mainly turn into the *non-cat* objects in the MBRNN model (see Figure 3.B.1).

Various cases can appear the *cat* samples in the models. When inspecting the *cat* samples in the MBRNN model, we find that some objects such as `SDSS J014904.18+243502.4`, `SDSS J104307.62+084059.2`, and `SDSS J101223.88+161313.4` might not have reliable photometric input features due to neighbor objects. Star-forming galaxies with emission lines are also found as the *cat* samples. For example, `SDSS J085139.46+455518.4` and `SDSS J091022.97+164534.9` have emission-lines representing star formation at redshifts of 0.28 and 0.30, respectively. Objects like `SDSS J150912.92+344418.1` at $z_{spec} = 0.06028$ have a large apparent size, and their photometry might not be reliable.

Table 3.C.2 Metrics for the ensemble cases with different redshift bins.

Number of bins	Bias	MAD	σ	σ_{68}	*NMAD*	R_{cat}
32	0.0015	0.0255	0.0392	0.0274	0.0260	0.0084
64	**0.0010**	**0.0253**	**0.0389**	**0.0272**	**0.0255**	0.0082
128	0.0014	0.0254	0.0390	0.0272	0.0256	**0.0082**
64 & 128	0.0013	0.0253	0.0389	0.0272	0.0255	0.0083

3.C Results With Different Ensemble Learning Configurations

Furthermore, we present the performance of the E3 ensemble model with 32 and 128 uniform redshift bins as well as the results obtained from the runs of the adopted 64 bins. Moreover, we consider merging the single models trained with 64 and 128 redshift bins rather than only the results with the same number of redshift bins. In particular, four single models sharing the same number of bins and trained with different anchor loss γs — 0, 0.2, 0.5, and 1 — are combined for the ensemble models of 32, 64, and 128 bins, respectively. For the merged ensemble model of 64 and 128 bins, however, eight single models are combined; four each of the 64 and 128 bin models trained with different γ values. Merging the models with the different number of redshift bins is conducted by interpolating their probability outputs to a higher number of redshift bins while maintaining the probability sum equal to 1.

Table 3.C.2 compares the point estimation metrics of the ensemble models with 32, 64, and 128 bins, and the case of combining models generated with 64 and 128 redshift bins. We find no advantage or improvement in these cases over the adopted ensemble model of combing the runs with 64 redshift bins.

Moreover, we check how the high γ values of the anchor loss affect the performance of the ensemble model. We attempt the E3 ensemble method with 64 uniform redshift bins while combining the additional results with γ in the order of 2, 5, and 8 into the set of models used in the adopted E3 model ($\mathcal{G}$), which is described in Section 3.4.2 as the model combining the results with $\gamma = 0, 0.2, 0.5$, and 1. Table 3.C.3 illustrates the metrics of the models for the trial configurations. Compared with the other runs, the ensemble model with the set $\mathcal{G}$ achieves the

Table 3.C.3 Metrics for ensembling cases with the different anchor loss γ values. The set $\mathcal{G}$ contains the models trained with anchor loss γ of 0, 0.2, 0.5, and 1.

Set of γ	Bias	MAD	σ	σ_{68}	*NMAD*	R_{cat}
$\mathcal{G}$	**0.0010**	**0.0253**	0.0389	**0.0272**	**0.0255**	**0.0082**
$\mathcal{G} \cup \{2\}$	0.0014	0.0253	0.0389	0.0272	0.0256	0.0083
$\mathcal{G} \cup \{2, 5\}$	0.0024	0.0255	**0.0389**	0.0274	0.0260	0.0082
$\mathcal{G} \cup \{2, 5, 8\}$	0.0040	0.0258	0.0389	0.0279	0.0268	0.0083

highest accuracy in general.

3.D Examination of The Ensemble Model Calibration

Many typical modern NNs are not well-calibrated (Guo et al. 2017). Hence, we examine the calibration of our E3 ensemble model. A well-calibrated model should have similar mean accuracy and confidence values. Figure 3.D.2 shows that our ensemble model E3 is well-calibrated. The difference between the mean accuracy and confidence of our E3 ensemble model is ~0.0078, which is close to 0. A reliability diagram depicting the accuracy variation as per the confidence confirms the adequate calibration of the ensemble model, almost identical to the completely calibrated case.

3.E Effects of The Galactic Extinction Correction

We apply the galactic extinction correction to the input colors using $E(B-V)$ and a simple correction rule, and then we train our model with corrected colors. Therefore, the model learns without $E(B-V)$ as a training input feature in this experiment. Three different correction rules are tested with the given $E(B-V)$, and the three different correction results (I, II, and III cases in Table 3.E.4) are compared in terms of model performances. Case I adopts the correction given as equations (7) to (13) in Tonry et al. (2012) with the observed apparent $(g-i)$ color. We note that this correction is fundamentally incorrect because the correction is valid only with

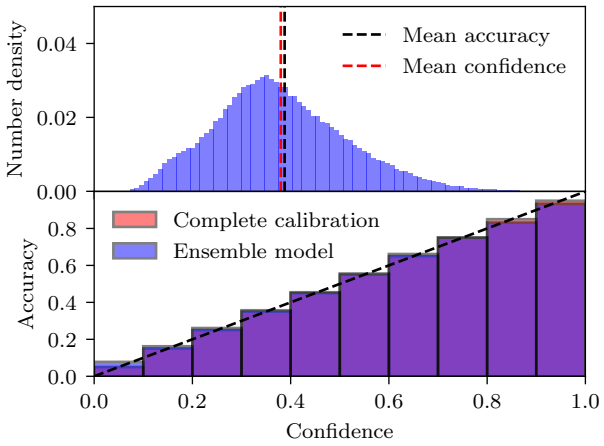


Figure 3.D.2 *Top*: Confidence distribution in the E3 model. Black and red vertical lines indicate the mean accuracy and confidence of the E3 model. *Bottom*: Reliability diagrams of the E3 model. Red and blue bars represent the distribution of the complete calibration and the E3 model result, respectively. The black dashed line marks a line with a slope of 1, meaning the perfect calibration.

the intrinsic color $(g - i)$, which we do not know beforhand, rather than the observed color (see Galametz et al. 2017, for discussions). The correction in cases II and III do not depend on $(g - i)$. Their correction rules adopt the representative values of the extinction A given as A at the pivot in Galametz et al. (2017) and A values used in Schindler et al. (2019) for cases II and III, respectively. Although the metrics in case II are the lowest overall, there are no significant differences among these three cases.

In case II, we analyze the effect of the correction on the photometric redshifts inferred by the trained model. We split $E(B - V)$ values into three different ranges, and each range has ∼33% of samples as low, middle, and high $E(B - V)$ values. Then, we compare photometric redshifts derived from our main model trained with $E(B - V)$ as an input feature (z_{model}) with case II results acquired with the extinction-corrected color data (z_{comp}). The left panel of Figure 3.E.3 shows that the redshift difference ($\Delta z = z_{model} - z_{comp}$) distribution. Interestingly, the distributions of redshifts for the low and high $E(B - V)$ values are positively and negatively shifted from a non-bias line ($\Delta z = 0$), respectively. This pattern is conspicuous in the right

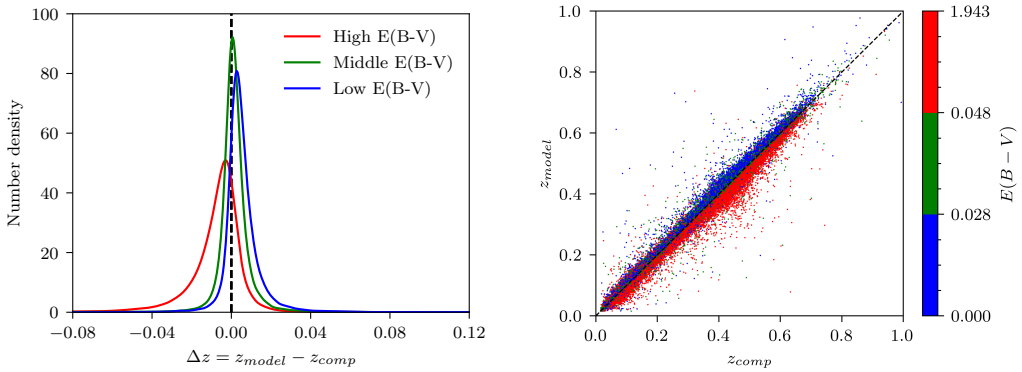


Figure 3.E.3 Distribution of the difference between photometric redshifts (z_{model}) from the model trained with $E(B-V)$ and those (z_{comp}) from the model trained without $E(B-V)$ but with the Galactic extinction-corrected colors (*left*). The samples are grouped into the low, medium, and high $E(B-V)$ value groups in which each group has approximately one-third of the number of samples. Δz shows systematic differences depending on the $E(B-V)$ values. This pattern is also found in the redshift comparison plot (*right*). The redshifts of the low and high $E(B-V)$ samples are mostly in the under- and over-estimation regions with respect to the correspondence line, which is represented by the dashed line, respectively.

panel of the figure, which shows a comparison between z_{comp} and z_{model}. These results indicate that the model trained with the extinction-corrected colors tends to under- and over-estimate redshifts compared to z_{model} for low and high $E(B-V)$ objects, respectively, indicating that the systematic bias is induced by the Galactic extinction correction on colors.

Figure 3.E.4 shows the distribution of the $z_{spec} - z_{model}$ and $z_{spec} - z_{comp}$, where z_{spec} corresponds to true spectroscopic redshifts. The distribution reported in the model with $E(B-V)$ as input features does not vary with respect to $E(B-V)$. However, the distribution of $z_{spec} - z_{comp}$ reveals that the trend found in Figure 3.E.3 can be explained by the fact that the extinction correction introduces bias into the redshift inference.

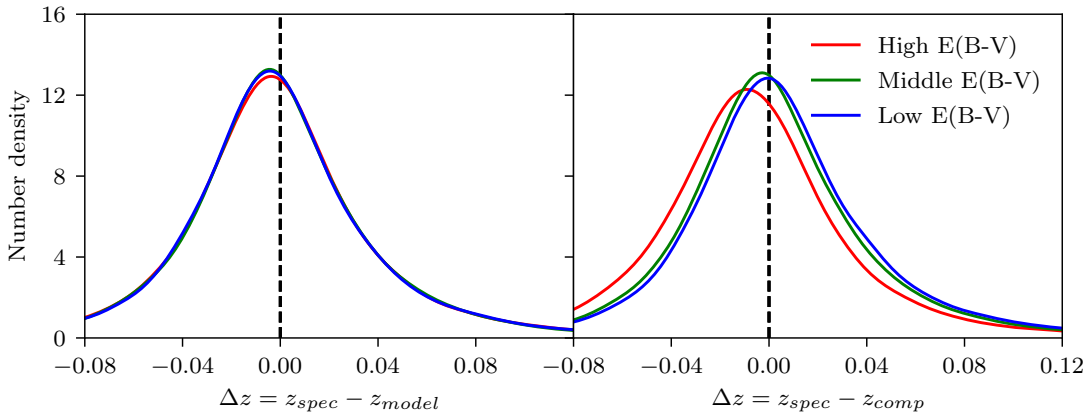


Figure 3.E.4 Distributions of the difference between spectroscopic redshift (z_{spec}) and photometric redshifts (z_{model} and z_{comp}). The samples are binned with respect to $E(B-V)$ as done for Figure 3.E.3. While the difference distributions for the z_{model} (*left*) are unbiased regarding $E(B-V)$, the distributions of the z_{comp} with higher $E(B-V)$ are shifted towards the lower Δz direction (*right*).

Table 3.E.4 Metrics for the experiments with three different Galactic extinction correction methods.

Case	Bias	MAD	σ	σ_{68}	*NMAD*	R_{cat}
I	**0.0019**	0.0261	**0.0403**	0.0280	0.0265	0.0090
II	0.0019	**0.0261**	0.0404	**0.0280**	**0.0264**	**0.0089**
III	0.0019	0.0262	0.0405	0.0281	0.0265	0.0092

Chapter 4

Estimation of Photometric Redshifts. II. Identification of Out-of-Distribution Data with Neural Networks [1]

Abstract

In this, the second paper of the series, we present a three-stage training strategy of neural networks (NNs) for both photometric redshift estimation of galaxies and detection of out-of-distribution (OOD) objects. Our strategy comprises supervised learning as well as unsupervised learning which enables the use of unlabeled (UL) data for OOD detection in training the NNs. Employing the UL data, which is the dataset most similar to the real-world data, ensures a reliable usage of the trained model in practice. We quantitatively assess the model performance of photometric redshift estimation and OOD detection using in-distribution (ID) galaxies and labeled OOD samples such as stars and galaxies. Our models produce well-matched photometric redshifts with spectroscopic redshifts for the ID samples and identify labeled OOD objects well

[1] To be submitted to the Astronomical Journal

Table 4.1 Spectroscopic QSO samples.

Dataset name	Number of objects	Selection conditions	Reference
SDSS DR15	290255	(CLASS == QSO) and (ZWARNING == 0) and (Z_ERR > 0.0)	Aguado et al. (2019)
LAMOST DR5	32793	(CLASS == QSO) and (Z > -9000)	Cui et al. (2012)
OzDES	772	(TYPES == AGN or QSO) and (FLAG ! = 3 and 6) and (Z >= 0.0025)	Childress et al. (2017)
PRIMUS	155	(CLASS == AGN) and (ZQUALITY == 4)	Cool et al. (2013)
COMOS-Magellan	53	(CLASS == bl or bnl or bal) and (Z_CONF == 4)	Trump et al. (2009)
6dFGS	49	(QUALITY_CODE == 6) or (REDSHIFT > 1.0)	D. H. Jones et al. (2009)
COSMOS-DEIMOS	30	(QF >= 10) and (Q >= 1.6)	Hasinger et al. (2018)
VVDS	16	ZFLAGS == 14 or 214	Le Fèvre et al. (2013)
COSMOS-Z-COSMOS	5	(CC == 14 or 214) and (REDSHIFT >= 0.0002)	Lilly et al. (2007, 2009)

with over 98% accuracy. Although quantitative assessment is impracticable due to the lack of labels and spectroscopic redshifts, we also find that the NNs successfully estimate reasonable photometric redshifts for ID-like UL samples and filter out OOD-like UL objects. The code for the model implementation is available at `https://github.com/GooLee0123/MBRNN_OOD`.

4.1 Data

As described in Chapter 3, we use the photometric data retrieved from the public data release 1 of Pan-STARRS1 (PS1) survey as input of the model (Kaiser et al. 2010). The PS1 survey provides photometry in five *grizy* bands (Chambers et al. 2016). The input data comprise seventeen color-related features: four colors $g-r$, $r-i$, $i-z$, and $z-y$, their uncertainties in PSF and Kron measurements, and reddening $E(B-V)$. The uncertainties are derived in the quadrature rule. We use only valid photometric data, which are found with the condition of *ObjectQualityFlags* == *QF_OBJ_GOOD* (Flewelling et al. 2020). In order to make each input feature contribute a

Table 4.2 Spectroscopic star samples.

Dataset name	Number of objects	Selection conditions	Reference
LAMOST DR5	4131528	(CLASS == STAR) and (Z > -9000)	Cui et al. (2012)
SDSS DR15	544028	(CLASS == STAR) and (ZWARNING == 0) and (Z_ERR > 0.0)	Aguado et al. (2019)
PRIMUS	1730	ZQUALITY == -1	Cool et al. (2013)
OzDES	1138	FLAG == 6	Childress et al. (2017)
COSMOS-DEIMOS	372	(REMARKS == STAR) and (Q >= 1.6)	Hasinger et al. (2018)
COSMOS-Z-COSMOS	300	REDSHIFT < 0.0002	Lilly et al. (2007, 2009)
C3R2-Keck	1	(REDSHIFT < 0.001) and (REDSHIFT_QUALITY == 4)	Masters et al. (2017, 2019)

similar amount of influence to a model loss function, we rescale input features using min-max normalization in a feature-wise manner.

We use three different sets of data in training and validating our model: ID, UL, and LOOD datasets. Each of the three datasets is used for different purposes as explained later. The ID data are galaxy samples used as training data for photometric redshift estimation. We use 1,480,262 galaxies, which are identicial to the samples used in Paper I, and assign 80%, 10%, and 10% of the samples to the training, validation, and test sets in estimating photometric redshifts. Because we randomly split data and have plenty of samples, it is reasonable to assume that the samples allocated to each set are drawn from the same distribution as the ID samples.

The UL data are unlabeled samples presumably containing both ID and OOD samples, and we do not have information about their physical classes and spectroscopic redshifts. We use the UL data for the unsupervised training of the model for the purpose of OOD detection. We construct the entire UL dataset to have 300,055,711 unknown objects. From the UL dataset, we evenly draw samples according to RA for unbiased training of the model regarding OOD detection. We use an RA interval of 10 degrees to divide the UL dataset and randomly choose 1,000,000 samples per RA interval. Hence, 36,000,000 UL samples are drawn from the entire UL dataset. Then, we assign 80% of the samples to the training set and each 10% of samples to the validation and test set.

The LOOD data consist of physically OOD objects which are physically different from galaxies. In this work, we define labeled (i.e., spectroscopically classified) QSOs and stars as the LOOD data. We obtain the photometric data of 324,234 QSOs and 4,681,989 stars in the PS1 data following the same selection condition adopted for the ID training samples. The sources of these spectroscopic objects are summarized in Tables 4.1 and 4.2. We use the LOOD data for the quantitative assessment of OOD detection performance and exclude it from training any models. Hence, we use the entire LOOD samples as the test data for the purpose of model validation.

4.2 Method

To equip the NNs with OOD scoring/detection functionality, we train two NNs F_1 and F_2 sharing the same structure with a multi-stage strategy composed of supervised and unsupervised steps (hereafter, $step_{sup}$ and $step_{unsup}$). The training strategy comprises three stages: supervised pre-training (training stage-1, TS1), iterative supervised and unsupervised training for OOD scoring/detection (TS2), and supervised training for photometric redshift estimation (TS3). While the first two stages of the training procedure are originally proposed in Yu & Aizawa (2019), the third stage is added to improve the model performance for the original task, i.e., photometric redshift estimation.

We adopt the supervised approach from Paper I for photometric redshift estimation as $step_{sup}$ and the unsupervised method introduced in Yu & Aizawa (2019) for OOD detection as $step_{unsup}$. In the following subsections, we outline each of $step_{sup}$, $step_{unsup}$, three-stage training procedure, and assessment metrics for model performance measures. For a more detailed explanation about each training step, refer to Paper I and Yu & Aizawa (2019).

4.2.1 Supervised Training Step for Photometric Redshift Estimation

For the $step_{sup}$, we consider discretizing redshift ranges and classifying samples into binned redshift intervals instead of performing photometric redshift regression, which we refer to as multiple-bin regression with the NN. For the method, we discretize the redshift range of train-

ing data and divide it into 64 independent bins with a consistent width. Then, the model estimates probabilities that the photometric redshifts of input samples lie in each bin. Using the model output probabilities, we may obtain point estimated photometric redshift z_{phot} by averaging central redshift values of the bins with the output probabilities.

As a training loss of the $step_{sup}$, we use anchor loss, L_{ANCH} (Ryou et al. 2019). The loss is designed to measure the difference between given two probability distributions considering prediction difficulties of individual samples caused by various reasons, e.g., the lack of data or the similarities between samples belonging to different classes. Anchor loss assigns the difficult samples for prediction high weights governed by a weighting parameter, γ. One can find the in-depth definition of the loss in Ryou et al. (2019). Since we train two networks, the training loss of the $step_{sup}$ is set as below.

$$L_{sup} = L_{ANCH}(p_1(\mathbf{z}|\mathbf{x_{ID}})) + L_{ANCH}(p_2(\mathbf{z}|\mathbf{x_{ID}})), \tag{4.1}$$

where p_1 and p_2 are the model output probability distributions coming from F_1 and F_2, respectively, $\mathbf{z}$ is a redshift bin vector, and $\mathbf{x_{ID}}$ is an ID input vector.

4.2.2 Unsupervised Training Step for Out-of-Distribution Detection

The $step_{unsup}$ uses UL data, assumably containing both ID and OOD samples, in training models and employs the disparity of the results from two networks to classify the samples as either ID or OOD. Identically structured networks supervisedly trained on the same ID data may have different results on OOD samples due to stochastic effects, although the networks are not optimized for OOD detection. Then, by defining discrepancy loss (L_{DCP}) to measure the disagreement and maximizing it, the networks can be forced to produce divergent results. The L_{DCP} defined for this purpose is subject to the following equation:

$$L_{DCP}(p_1,\ p_2) = H(p_1(\mathbf{z}|\mathbf{x})) - H(p_2(\mathbf{z}|\mathbf{x})), \tag{4.2}$$

where $H(\cdot)$ is the entropy. As the networks are trained to maximize the loss, the entropies of F_1 and F_2 outputs respectively increase and decrease. Namely, F_1 outputs a flat probability distribution, and F_2 generates a peaked one. Note that the loss is available for the unsupervised approach using UL samples since it can be computed without sample labels.

To prevent divergent results on ID samples, we employ the sum of L_{sup} and L_{DCP} as the training loss of $step_{unsup}$. Although the unsupervised approach using L_{DCP} diverge the model outputs on OOD examples, it also can cause discrepant results on ID samples since UL data includes ID as well as OOD samples. Hence, defining the training loss of $step_{unsup}$ as such may help the two model outputs stay as similar as possible. The training loss of $step_{unsup}$, L_{unsup}, is defined as below.

$$L_{unsup} = L_{sup} + L_{DCP}(p_1(\mathbf{z}|\mathbf{x_{UL}}),\ p_2(\mathbf{z}|\mathbf{x_{UL}})), \tag{4.3}$$

where $\mathbf{x_{UL}}$ is an UL input vector.

4.2.3 Three-Stage Training

We proceed with three-stage training of the networks using the aforementioned $step_{sup}$ and $step_{unsup}$, as shown in Figure 4.1. TS1 is a pre-training of the networks with the $step_{sup}$. The purpose of this stage is to obtain a stable performance of the models for photometric redshift estimation. In the TS2, the networks are trained for OOD detection using the iterative training step, comprised of one $step_{sup}$ and two $step_{unsup}$. The iterative step precludes the divergence of the two network outputs on ID samples as the L_{sup} added to the L_{unsup} does. Hence, as the training continues, the disagreement of the model outcomes on OOD samples becomes larger than ID samples since the OOD samples are outside the support of L_{sup} and $step_{sup}$, which are only applied to ID galaxies. Then, the measure of the disagreement, L_{DCP}, can be used as an OOD score to flag OOD candidates. At the inference level, the trained networks at TS2 are used to score the test samples as OOD.

In addition to the two-stage training, we proceed to additional supervised training on ID samples for photometric redshift estimation, i.e., TS3. The unsupervised training for OOD detection infects the model performance of the original task since the model is optimized for two different losses. Training a single model performing multiple tasks with greater or equal performance compared to a uni-task model is out of the scope of this paper. Therefore, we train the two networks again using the L_{sup} after TS2. Then, the outputs of the two TS3 networks are combined by averaging to estimate photometric redshifts.

Lastly, we stress that, at TS3, the networks can be trained using the L_{sup} with different γs of

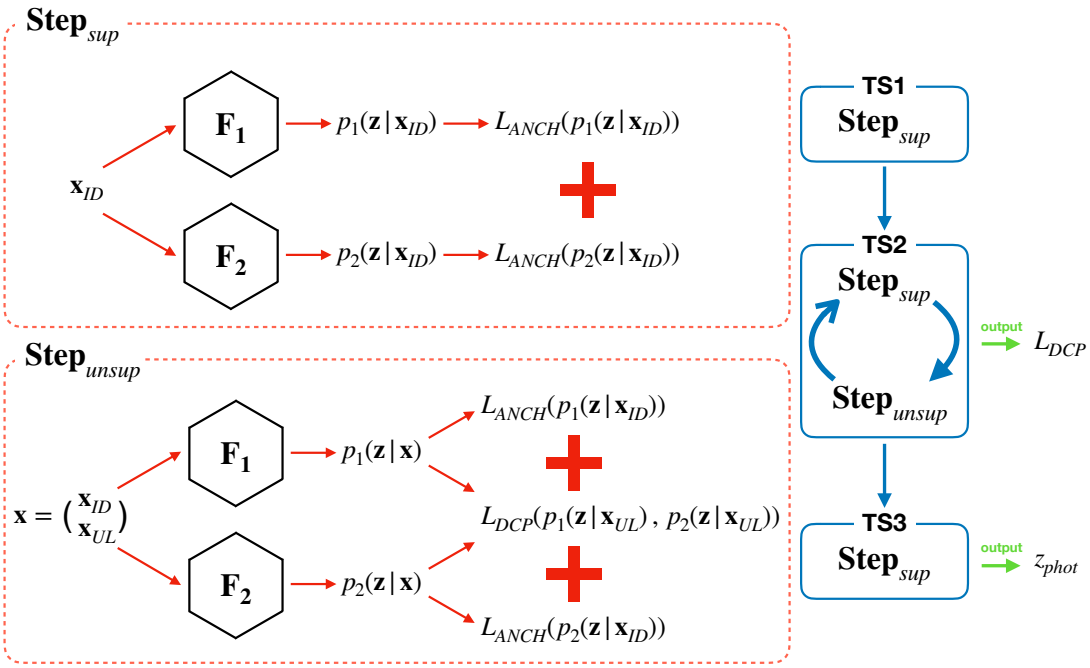


Figure 4.1 Schematic image of three-stage training. The subscripts ID and UL of **x** indicate that the **x** belongs to the ID and UL dataset. The training stages comprise two learning steps: supervised step ($step_{sup}$, *upper red box*) and unsupervised step ($step_{unsup}$, *bottom red box*). The $step_{sup}$ is designed for photometric redshift estimation, and $step_{unsup}$ is contrived for OOD scoring/detection. Note that the type of loss for each step is different. As the first step of the entire training procedure (training stage-1, TS1), the networks are pre-trained with $step_{sup}$. The pre-trained networks, then, are transmitted to the TS2 and trained for OOD scoring/ detection using iterative $step_{sup}$ and $step_{unsup}$. In the last step, TS3, the networks are again optimized for photometric redshift estimation only using $step_{sup}$.

L_{ANCH} and integrated through ensemble learning. In this manner, we may reduce the bias of the single models caused by the anchor losses in L_{sup}. However, in this paper, we only introduce the results of the single model trained with γ of 0 since we have already demonstrated the improved performance of the ensemble model in Paper I.

4.2.4 Assessment Metrics

Since our networks perform redshift point estimation and OOD detection, we need independent metrics to assess the network performances on both tasks. As point estimation metrics, we adopt the ones used in Paper I: *Bias, MAR,* σ*,* σ_{68}*, NMAD, and* R_{cat}. A detailed explanation about the metrics is offered in Paper I. Note that the lower the point estimation metrics are, the higher the quality of the redshifts is.

Prior to explaining OOD detection metrics, we first define four measures yielding the metrics. *True negative (TN)*, *true positive (TP)*, *false negative (FN)*, and *false positive (FP)* are all cases to which the classified samples belong either. The boolean values (*true* and *false*) of the terms indicate if the predicted class of the sample is correct or incorrect. The negative and positive mean the actual class in which the given sample is included. Here, we respectively set ID and OOD as negative and positive since the networks perform OOD detection. Hence, for example, true positive is defined as the number of correctly classified OOD samples in our case. Note that the values can be varied with respect to the threshold of the OOD score. We use the central value of the minimum and maximum L_{DCP} as the threshold to compute the metrics. The following is brief explanations of the detection metrics:

- *Accuracy*: ratio of the correctly classified samples to the entire samples $\frac{TN+TP}{TN+TP+FN+FP}$,
- *Precision*: a metric quantifying the ratio of the correctly classified positive samples to the entire positively classified samples $\frac{TP}{TP+FP}$,
- *True Positive Rate (TPR)* or *Recall*: a fraction between the correctly predicted positive samples and the entire positive samples $\frac{TP}{TP+FN}$,
- F_β: a metric measuring harmonic mean of precision and recall weighted by β $\frac{(1+\beta^2)*Precision*Recall}{\beta^2*Precision+Recall}$. The lesser and larger β weight more on precision and recall, respectively. Typical values

for β is 0.5, 1.0, and 1.5. In this work, we set $\beta = 1.5$ ($F_{1.5}$) since we want to more focus on the recall than precision, i.e., the number of the correctly classified OOD samples out of all OOD samples,

- AUC_{ROC}: Area Under the Receiver Operating Characteristic Curve (ROC curve). ROC curve is a visualized measure of the detection performance of the binary classifier. In the curve, the *TPR* is plotted in function of the false positive rate for different thresholds of the OOD score. False Positive Rate is a fraction between the incorrectly predicted negative samples and the entire negative samples $\frac{FP}{TN+FP}$. AUC_{ROC} offers a comprehensive measure of the detection performance of the network,

- AUC_{PR}: Area Under the Precision-Recall Curve (PR curve). Although AUC_{ROC} offers a representative measure of the binary detection performance for most of the cases, it may cause misleading results for imbalanced data. It is because *TPR* (y-axis of the ROC curve) only depends on positives, so it excludes the consideration of the negative class. In the PR curve, precision is plotted in the function of recall. Because precision takes positive and negative samples into account, AUC_{PR} may be more informative for skewed data.

In contrast with point estimation metrics, higher detection metrics indicate better detection performance of the networks.

4.3 Results

In this section, we present the results of the model testing for photometric redshift estimation and OOD detection. We emphasize that all the results in this section are produced with the samples in test set, unless we state otherwise.

4.3.1 Photometric Redshift Estimation on In-Distribution Samples

We find that the TS3 networks outperform TS2 networks when it comes to photometric redshift estimation. To vindicate the additional stage-3 training of the networks, we compare the

Table 4.3 The metrics for photometric redshift estimation of the models from TS2 and TS3.

Variables	metrics					
Case	Bias	MAR	σ	σ_{68}	*NMAD*	R_{cat}
TS2 - HE	0.0306	0.0466	0.0959	0.0343	0.0314	0.0637
TS2 - LE	**0.0015**	0.0281	0.0444	0.0293	0.0280	0.0150
TS3 - avg	0.0020	**0.0254**	**0.0392**	**0.0272**	**0.0256**	**0.0087**

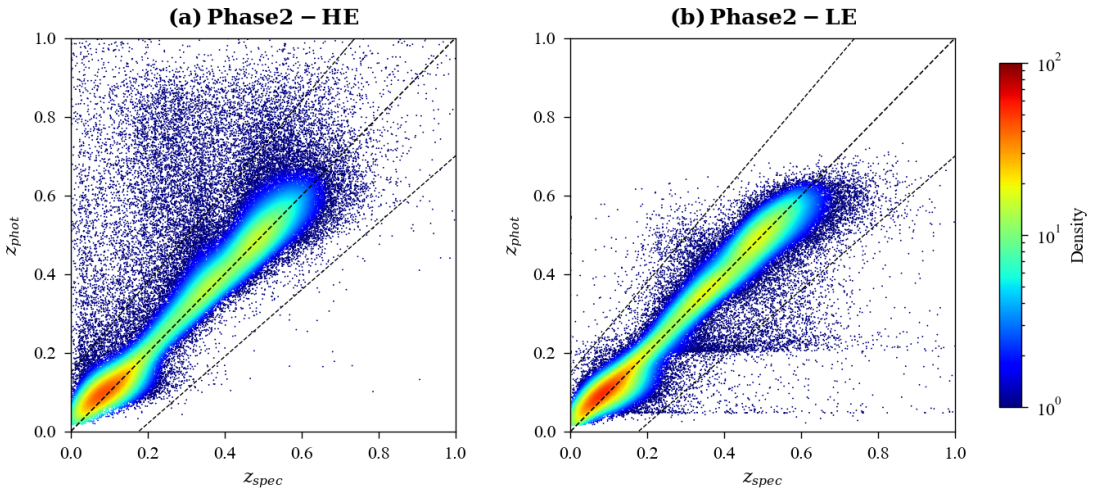


Figure 4.2 Comparison between spectroscopic and photometric redshifts from HE (*left*) and LE (*right*) models at TP2. The scatters are color-coded according to density. The dashed lines at the center and both sides of each plot are slope-one lines and catastrophic error boundaries, respectively.

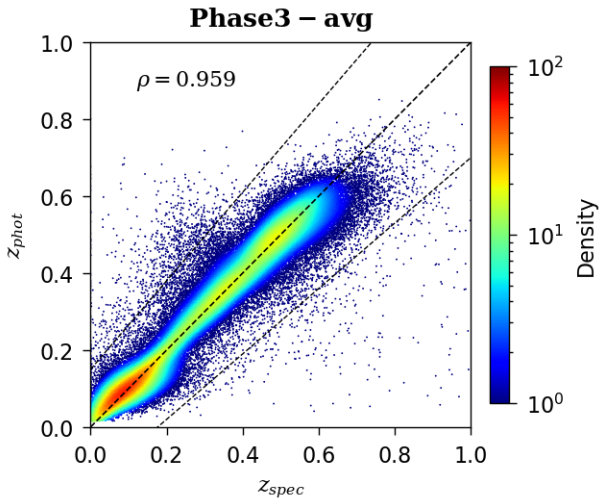


Figure 4.3 Comparison between spectroscopic and photometric redshifts from the averaged network at TS3, color-coded with density. The Pearson correlation coefficient is written on the upper left corner of the plot. Note that the redshift overestimation or collapse found in TS2 model disapears for the TS3 model.

photometric redshift qualities from TS2 and TS3 models. Table 4.3 shows the point estimation metrics from the high and low entropy models (HE and LE, formerly F_1 and F_2) at TS2 and averaged model at TS3. Except for the bias, the network at TS3 outperforms the other networks. It proves that unsupervised training of the networks at TS2 for OOD detection infects the photometric redshift qualities from the networks.

The networks at TS2 hold noticeable peculiarities deteriorating photometric redshift qualities. Figure 4.2 displays the comparison between spectroscopic and photometric redshifts from the HE and LE models at TS2. The photometric redshifts from the HE model are partially overestimated. On the other hand, the photometric redshifts from the LE network are collapsed at $z_{phot} \sim 0.05$ and $z_{phot} \sim 0.2$. The peculiarities of estimated redshifts arise from the maximization of L_{unsup} during TS2, which makes the output probability distributions of HE and LE models flat and peaked. In addition to that, the density peaks of both networks are misaligned with the slope-one line, the low redshift regions of the density peaks being shifted upward. We believe that the strangely low bias of the LE model stems from the cancelation of the deviation

Table 4.4 The OOD flagging metrics of the networks on LOOD samples. Every metric of TS2 networks higher than 0.98 shows that the networks accurately detect OOD objects.

Variables	metrics					
Case	Accuracy	Precision	TPR (Recall)	$F_{1.5}$	AUC_{ROC}	AUC_{PR}
TS2	**0.9802**	0.9987	**0.9808**	**0.9863**	**0.9938**	**0.9998**
TS3	0.0295	**0.9990**	0.0008	0.0012	0.7547	0.9914

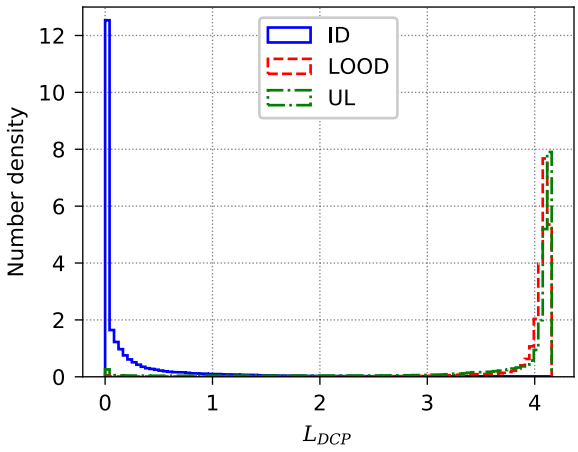


Figure 4.4 L_{DCP} distribution of ID, LOOD, and UL samples.

between the overestimated redshifts in the density peak and underestimated redshifts in the regions with collapsed redshifts.

These oddities vanish for the averaged network at TS3, as shown in Figure 4.3. In addition to that, the density peak of the samples is also well-aligned with the slope-one line. The Pearson correlation coefficient $\rho = 0.959$ affirms that the network estimated redshifts are well-matched with spectroscopic redshifts. Providing the peculiar behaviors of TS2 estimated redshifts and the TS3 network's low point estimation metrics overall, we argue that the networks at TS2 are nonoptimal for photometric redshift estimation and the necessity of the TS3 is justified.

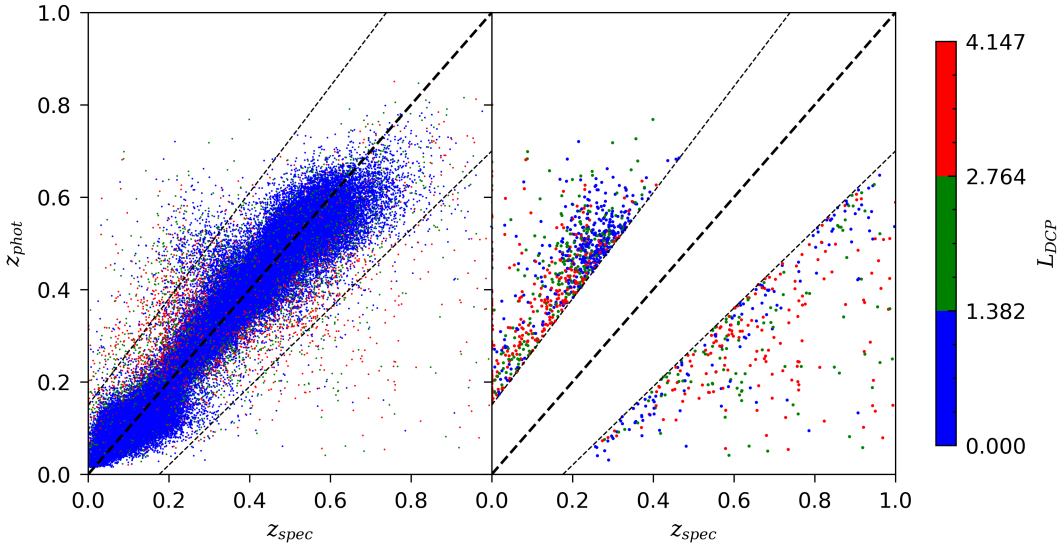


Figure 4.5 Distribution of L_{DCP} from TS2 networks in the space of spectroscopic and photometric redshifts from TS3 models. The *left* panel is for the entire test samples and the *right* panel is for the *cat* samples. For clear visualization, we divide L_{DCP} range into three regions with uniform width: low, middle, and high L_{DCP} ranges. Note that the samples with high L_{DCP} reside near or outside the catastrophic boundaries.

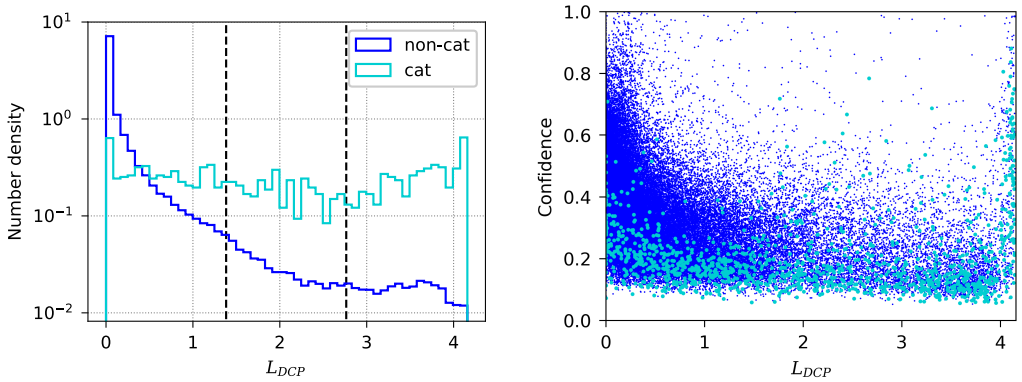


Figure 4.6 *Left*: Log-scale L_{DCP} distributions of ID *non-cat* (blue) and *cat* (azure) samples. Vertical dashed lines are boundaries of low, middle, and high L_{DCP} ranges. *Right*: Distribution of ID *cat* and *non-cat* samples in the space of L_{DCP} and confidence of TS3 networks with the same color scheme. For clear visualization, we use a larger marker size for the *cat* samples and plot *cat* samples on top of the *non-cat* ones.

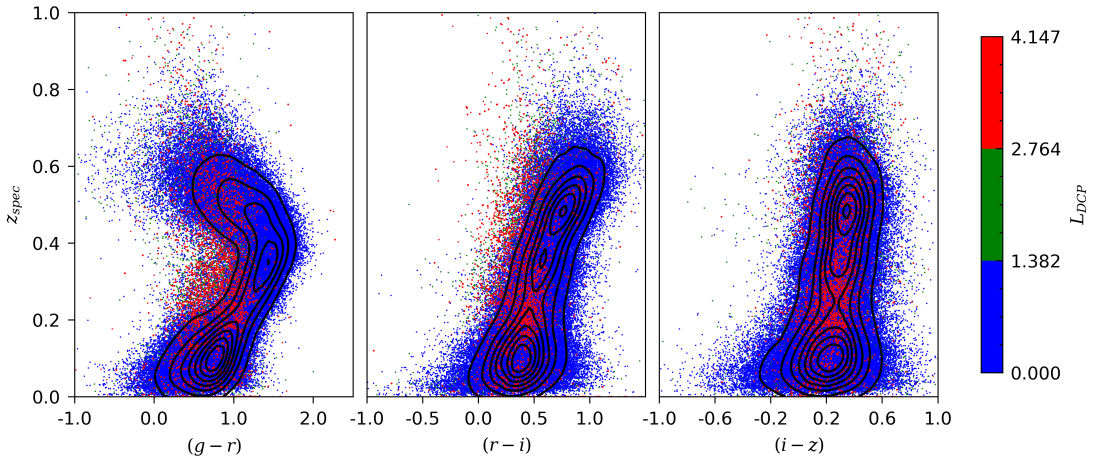


Figure 4.7 The L_{DCP} distributions of ID samples in the space of colors and spectroscopic redshifts. Contour lines depict the distributions of the training samples in each of the input spaces. Note that high L_{DCP} samples mostly lie outside the contour lines or low density regions.

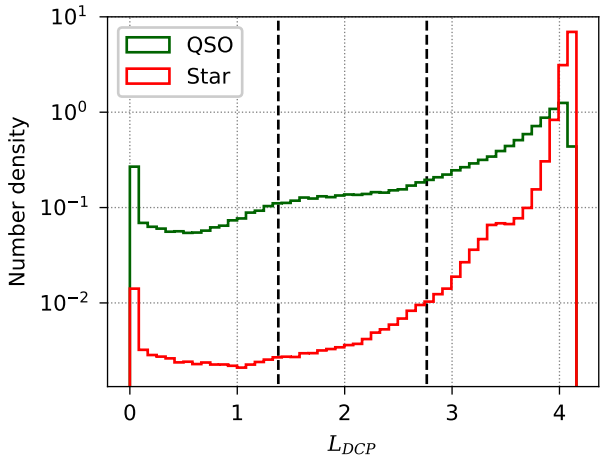
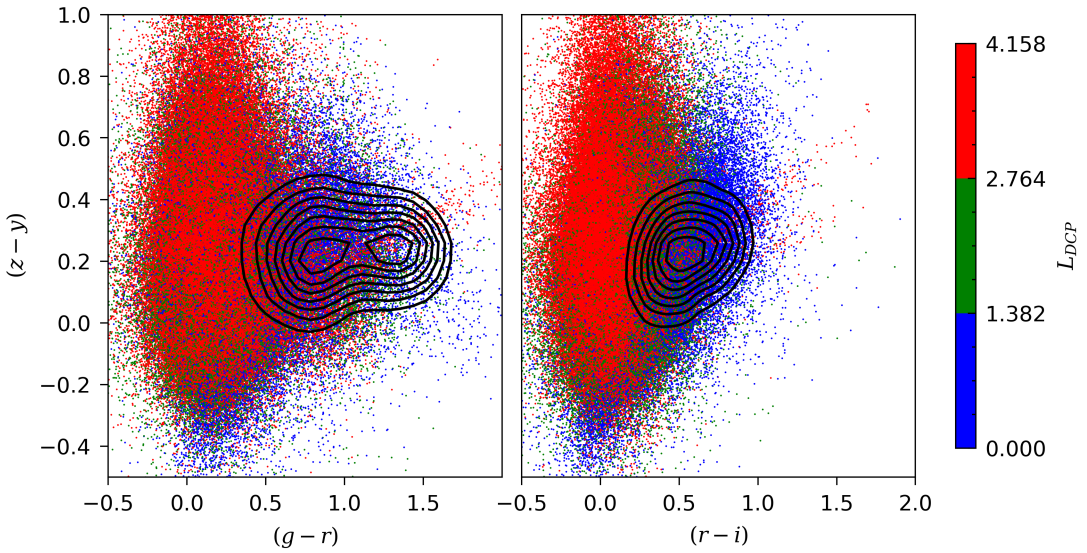
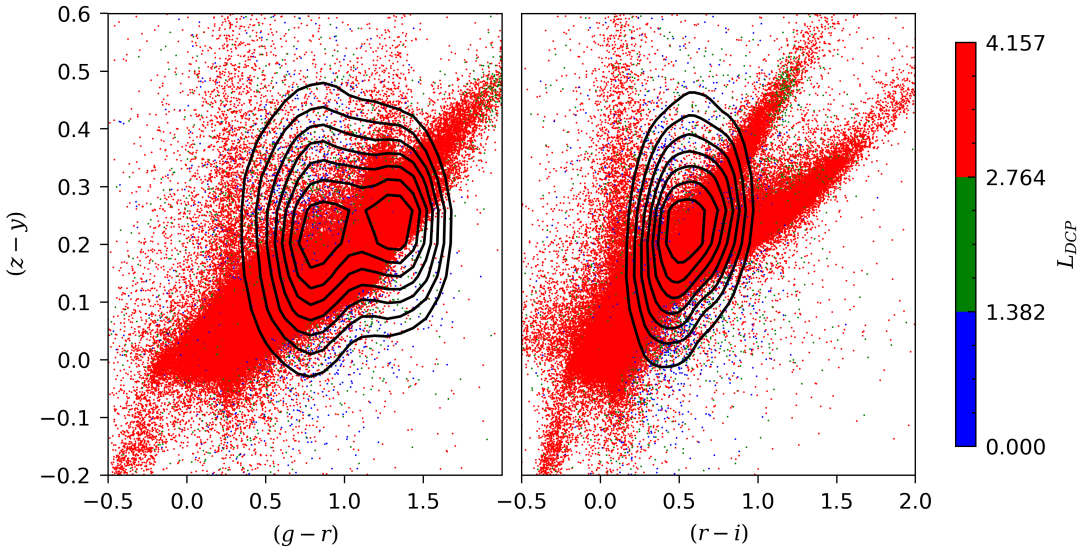


Figure 4.8 Log-scale L_{DCP} distributions of LOOD samples. The distributions of QSOs (dark-green) and stars (red) are drawn indepedently. Vertical dashed lines are boundaries of low, middle, and high L_{DCP} regions.



(a) QSO



(b) Star

Figure 4.9 The L_{DCP} distributions of QSOs (*upper*) and stars (*bottom*) in the space of input colors. The black contour lines depict the distribution of ID samples in the corresponding input spaces. The scatters are discretely color-coded according to L_{DCP} groups. The star samples used to plot the scattergram are randomly undersampled by a factor of about 10 from the entire star dataset.

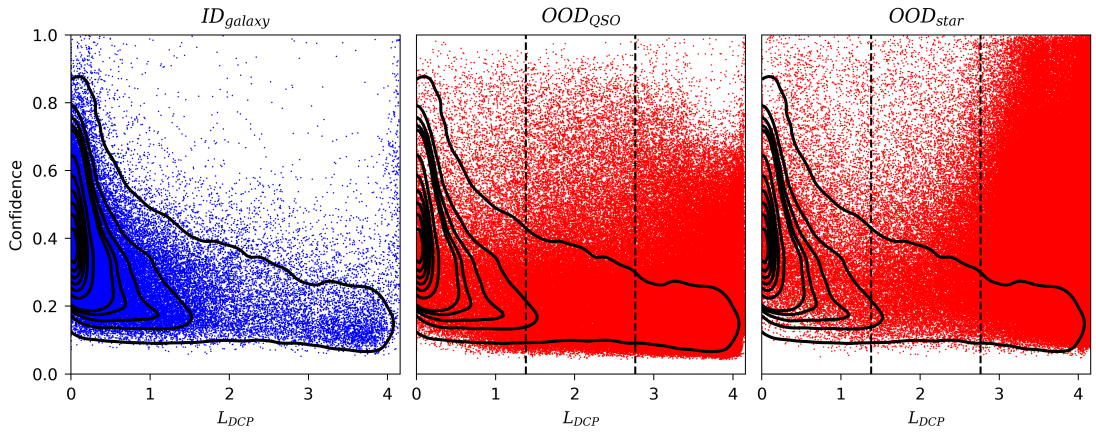


Figure 4.10 Distribution of galaxy (ID, *left*), QSO (OOD, *middle*), and star (OOD, *right*) samples in the two-dimensional space of L_{DCP} and confidence. Green contour lines indicate the *residence* of ID samples in the space. The guide values of L_{DCP} dividing low, middle, and high L_{DCP} regions are marked as vertical dashed lines in the panels of OOD samples.

4.3.2 Out-of-Distribution Score for Labeled Data

In contrast with photometric redshift estimation, the networks at TS2 show a better separation of ID and OOD samples than the TS3 networks. As mentioned in Section 4.2.2, multiple networks not optimized for OOD detection output different results on OOD samples. Hence, we compare the OOD detection performance of the networks at TS2 and TS3. The detection metrics of the networks at TS2 and TS3 are tabulated in Table 4.4. Apparently, the TS2 networks outperform the TS3 models as for OOD detection.

However, interestingly, the TS3 networks have higher precision than that of TS2 networks. It is due to high *TP* caused by the small number of samples flagged as OOD by TS3 networks; only 352,055 samples ($\sim 7\%$) amongst 5,006,223 OODs are correctly classified as OOD. This small number results in the unexpectedly high precision, AUC_{ROC}, and AUC_{PR} of TS3 networks. Providing the high precision of TS3 networks is overestimated, and TS2 networks outperform TS3 networks for all the other metrics, we use TS2 networks for OOD scoring and detection.

The TS2 networks well-separate ID and OOD samples, as shown in Figure 4.4. The figure

displays the L_{DCP} distributions of ID, LOOD, and UL samples yielded from the TS2 networks. While most of the ID samples are clustered in the low L_{DCP} ranges, the *residence* of the LOOD samples is in the vicinity of the highest L_{DCP}. In addition, the vast majority of the UL samples are distributed in the high L_{DCP} regions[2]. Providing this, we infer that the most fraction of UL samples are more likely to be OODs than IDs, although we cannot specify the classes of the samples. In the rest of this subsection, we focus more on the labeled data and defer the inspection of the UL samples to the following subsection.

We find that the prediction difficulties of photometric redshift estimation and OOD score for ID data are somewhat correlated. Figure 4.5 shows the L_{DCP} distribution in the space of the spectroscopic and photometric redshifts for ID samples. While most of the samples in the *non-cat* region are assigned to the low L_{DCP} group, many of the samples in the *cat* region belong to middle or high L_{DCP} groups. Besides, high L_{DCP} samples in the *non-cat* region deviate more from the slope-one line than low L_{DCP} samples. It indicates that the high OOD score samples, which can be viewed as OOD-like ID samples, are likely to have high error of photometric redshifts.

The log-scale L_{DCP} distributions of ID *non-cat* and *cat* samples in the left panel of Figure 4.6 also endorse our interpretation of the correlation. The distributions show that a higher ratio of *cat* samples populates in the high L_{DCP} ranges than *non-cat* samples. While the number of *non-cat* samples almost monotonically decreases as the L_{DCP} increases, the *cat* samples are almost uniformly distributed for overall L_{DCP}. Providing these consistent results, we deduce that the high prediction difficulties of redshift estimation and OOD-like ID samples may share the same origin[3].

We also find that TS3 networks are unreliable concerning high L_{DCP} *cat* samples. The right panel of Figure 4.6 shows the distribution of *non-cat* and *cat* samples in the space of the L_{DCP} and the confidence. As shown in the figure, while the number of high confidence *non-cat* samples decreases as L_{DCP} increases, *cat* samples mostly dwell in the low confidence ranges

[2] Although it is not conspicuous, there is a small peak of UL sample distribution at the lowest L_{DCP} bin. We more minutely investigate this in Section 4.3.3

[3] In Paper I, we have already found that the lack of training samples in the input space causes high errors of photometric redshifts.

throughout low and middle L_{DCP} ranges. It shows that the TS3 networks are well-calibrated about the results of low and middle L_{DCP} *cat* samples, which makes the model outcomes on low L_{DCP} samples more credential since the models show a gradually less confident attitude toward high-error samples. However, notice that the *residence* of high L_{DCP} *cat* samples extends to the high confidence regions, which is the indication of overconfident results of the TS3 networks for OOD-like ID *cat* samples. It again emphasizes the importance of the OOD detection since we can filter out these high OOD score samples that the model is overconfident about with catastrophic error.

To investigate the causes of OOD-like ID samples, we visually inspect the distribution of L_{DCP} in the space of input colors and spectroscopic redshift in Figure 4.7[4]. As shown in the figure, most of the high L_{DCP} samples reside outside the *residence* of the training samples. It suggests that the networks at TS2 are optimized towards high-contribution data during training and assign high OOD scores to the test samples similar to the low-contribution samples during training, i.e., these samples are low-contribution OODs in the view of the model. Providing that the low-contribution OODs cause high errors of photometric redshifts as presented in Paper I and the OOD-like samples are also distributed in the low-density regions, both redshift prediction difficulty and high OOD score regarding ID samples originate from the lack of training samples in the input space.

As ID samples with high L_{DCP} exist, there are low L_{DCP} LOOD samples, which can be thought of as ID-like OOD objects. Figure 4.8 depicts the log-scale L_{DCP} distribution of the two types of LOOD, i.e., QSOs and stars. As shown in the figure, a fraction of LOOD samples dwells in low and middle L_{DCP} ranges, although most of the samples bring high L_{DCP}. The points to be noted are 1) both QSO and star distributions have peaks at the lowest L_{DCP} bin, 2) the ratio of the lowest L_{DCP} peak to the highest L_{DPC} peak of QSO is larger than that of star, and 3) a higher ratio of QSOs is distributed in the low and middle L_{DCP} ranges than stars.

One of the main causes of these ID-like OOD samples is that they locate in the vicinity of the *residence* of the ID samples in the input space. Figure 4.9 shows the L_{DCP} distribution of

[4]In this visual expression, often used hereafter, displaying the projected dimensions of the input space, degeneracy concerning the input features which does not exist in the higher dimension of the space may emerge.

LOOD samples in the space of input colors. As can be seen in the upper panel of the figure, the high L_{DCP} QSO samples dominate low L_{DCP} ones outside the *residence* of ID samples. However, the number of the low L_{DCP} samples increases as approaching the density peak of the ID samples. It indicates that the QSOs holding similar input feature values with those of training data can be look-alike with ID samples to the trained model. These ID-like QSOs contribute to the peak at the lowest L_{DCP} bin of Figure 4.8.

On the other hand, these patterns are not noticeable for stars (see the bottom panel of Figure 4.9), although some of the objects certainly reside inside the ID *residence* and low L_{DCP} stars are certainly exist with a peak at the lowest OOD score bin (see Figure 4.8). We suppose that more stars than QSOs can be sorted from the ID samples in higher dimensions of input spaces since the distribution patterns of the stars are significantly different from those of ID samples. Hence, the relatively smaller peak of the stars at the lowest OOD score and a lower fraction of ID-like stars than QSOs arise from the distribution pattern difference.

Evidently, filtering out these low L_{DCP} QSO and star samples enables further reliable usage of the trained model. We can partially detach low and middle L_{DCP} LOOD objects residing outside the *residence* of ID samples in the two-dimensional space of L_{DCP} and confidence. As already seen in the right panel of Figure 4.6, the *residence* of ID samples with respect to the confidence gets narrower towards low confidence region as L_{DCP} increases. In Figure 4.10, it is clearly depicted by the density contour lines of ID samples. The figure presents the distributions of galaxy (ID), QSO (OOD), and star (OOD) samples. It is noticeable that many of the LOOD samples are distributed outside the ID contour lines. Hence, a considerable number of LOOD samples in the low and middle L_{DCP} range and outside the ID *residence* can be differentiated from the ID samples. On the other side, OOD-like ID samples cannot be distinguished from OOD because the *residences* of high L_{DCP} ID and LOOD samples are overlapped.

4.3.3 Out-of-Distribution Score and Photometric Redshifts of Unlabeled Data

Since the UL data have no labels and spectroscopic redshifts, we ought to assess the model performance on UL samples in an indirect manner. To measure the OOD detection performance, we first compare our results to the point-source score (ps-score, Tachibana & Miller 2018).

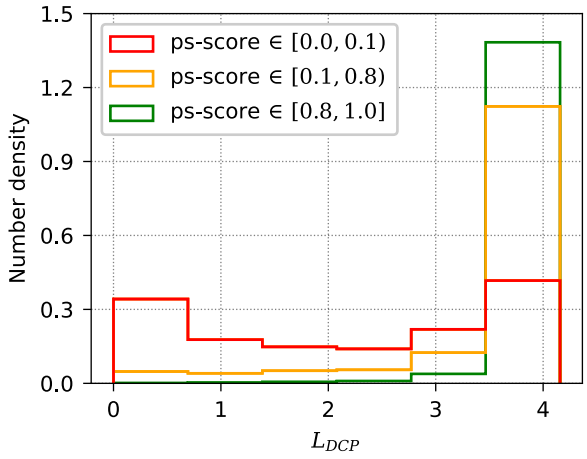


Figure 4.11 Distribution of L_{DCP} with respect to different ps-score range.

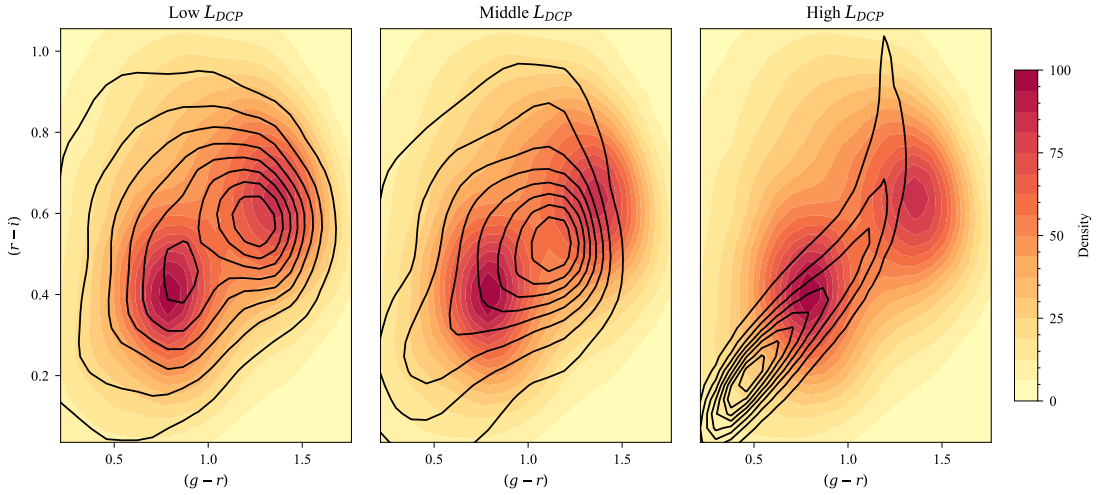


Figure 4.12 Density contour lines of the UL samples with filled density contour map of ID samples in the space of the input colors, $(g-r)$ and $(r-i)$. We again use the trisection of L_{DCP} to split UL samples into low, middle, and high L_{DCP} groups.

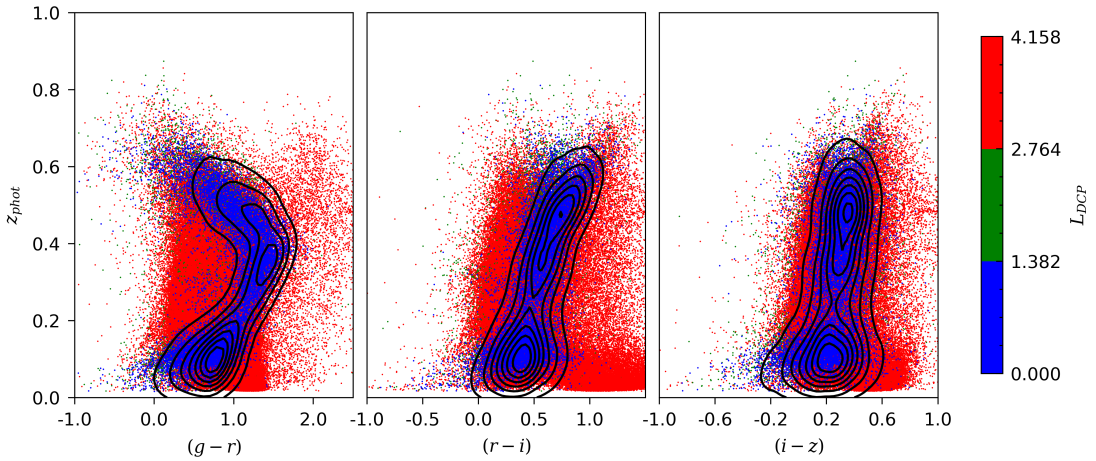


Figure 4.13 L_{DCP} distributions of UL samples in the space of input colors and photometric redshifts, color-coded with respect to L_{DCP} groups. Remember that the samples lack spectroscopic redshifts. The black contour lines are drawn with the spectroscopic redshifts and corresponding input colors of ID samples.

The ps-score is also a machine-learning-based probabilistic measure of sources of being point sources for PS1 objects in the range of [0, 1]. The ps-score outperforms a simple cut of (PSF − Kron) magnitudes for separating galaxies from stars. Although it is not always the case, the samples with a lower ps-score are more likely to be galaxies, and we expect our model to produce lower L_{DCP} for low ps-score samples and vice versa.

The distributions of L_{DCP} with respect to ps-score are subject to the expected pattern, although the two machine learning outputs are not perfectly matching. Figure 4.11 shows the L_{DCP} distribution of UL samples, appertaining to different ps-score ranges. The test examples with ps-score used for this examination are undersampled from the UL dataset within the RA range [0, 10], [100, 110], and [170, 180]. As shown in the figure, the samples within the lowest ps-score interval are mostly distributed near the minimum and maximum L_{DCP} values, and the distributions of the higher ps-score samples are shifted towards high L_{DCP} ranges. We conjecture that the middle/high L_{DCP} samples within the lowest ps-score are highly liable to be the low-contribution OOD galaxies residing in outskirts or low-density parts of ID *residence*, as mentioned in Section 3.1. Except for this understandable difference, the two machine learning

models reach a consensus on other UL samples.

In addition to the pattern shifts according to the ps-score, we find the analogous pattern variations regarding L_{DCP} in Figure 4.12, which also indirectly proves L_{DCP} to be a confidential OOD score. The figure displays how the *residences* of UL samples (contour lines) vary with respect to L_{DCP} compared to the ID sample *residence* (filled contours) in the space of input colors. The low L_{DCP} UL samples show the most resembling distributions to those of the ID samples. Viz, the networks assign low L_{DCP} to the samples with comparable input features to the ID samples. The middle L_{DCP} UL samples also show a similar form of *residence* to that of ID samples, although they do not reproduce the multiple peaks of ID samples. Evidently, the samples within the high L_{DCP} group have a completely different distribution from ID samples. Scilicet, in the increasing order of the OOD score, the density distributions of the UL data deviate more from those of ID samples. It implies that our trained networks successfully sort out OOD-like samples from ID-resemble ones.

We reckon that the networks also adequately produces photometric redshifts for UL samples. We expect that the distributions of photometric redshifts to be similar to that of ID samples for low L_{DCP} UL samples and to random guessing, i.e., uniform distribution, for high L_{DCP} UL samples. Figure 4.13 depicts the L_{DCP} distributions of UL samples in the space of three input colors and photometric redshifts. As shown in the figure, the samples within low L_{DCP} ranges mostly reside inside the *residence* of ID samples. It signifies that the photometric redshifts for ID-like UL samples are properly computed by the model according to input feature values, although we cannot quantify the error of the estimated redshifts due to the lack of spectroscopic redshifts. On the other hand, the photometric redshifts of OOD-like samples are randomly spread in the spaces without any patterns. It certainly shows that the photometric redshift distribution for OOD-like samples are more closed to uniform distribution than ID-like samples. However, most of the redshifts are in the region where $z_{phot} < 0.8$. Given that the maximum redshift value of the training set is approximately 2.00, it indicates that the individual model-estimated probability distributions for OOD-like samples are truncated at $z \sim 0.8$. It arises from the fact that about 99.82% of training samples have below 0.8 spectroscopic redshifts. During training, the networks are optimized to estimate probability distributions of redshifts shifted

towards the range where the majority of training spectroscopic redshifts are populated.

4.4 Discussion and Conclusion

Our model presented here successfully estimate photometric redshifts of galaxies even with the addition of the extra NN to our previous model presented in Paper I. The proposed multi-step approach in this Paper II reproduces a reliable estimation of photometric redshifts. Our trained model is available online[5] for further usages.

The proposed method to detect the OOD data with the photometric redshift inference model shows that the new model can definitely measure how much given data deviate from the training data as OOD samples when they are physically galaxies with the OOD properties. The majority of samples with the catastrophic redshift estimation corresponds to the data with high OOD scores (see Figure 4.5). However, our model estimating the OOD score cannot replace the model classifying out galaxies from stars and QSOs. As shown in Figure 4.10, a large number of stars and QSOs show low OOD scores and high confidence values for photometric redshifts in our model although they are definitely physically OOD objects. These objects have the indistinguishable input features compared to the real galaxies in terms of the model's perspective for photometric redshifts.

We plan to apply SED-fitting methods to the galaxies with the high OOD scores or low confidence values. The machine learning inference model has limitation depending on training samples. Therefore, using the SED-fitting methods as well as the machine learning inference models can be an effective way to achieve high analysis speed and accuracy together with broad applicability in estimating photometric redshifts for a large number of galaxies expected in the future surveys such as the Legacy Survey of Space and Time (Ivezić et al. 2019).

The current implementation including the two NNs needs to be improved if the full benefits of ensemble learning with the multiple anchor loss parameters presented in Paper I are required for better estimation of photometric redshifts and evaluation of the OOD scores in a single machine learning framework. The current model presented in this paper excludes multiple inferences of photometric redshifts in ensemble learning. The ensemble model distillation

[5] `https://github.com/GooLee0123/MBRNN_OOD`

might be a possible way to accommodate the multiple models included in the ensemble learning as a single combined model (e.g., Malinin et al. 2020) in the current framework combining inference of both photometric redshifts and OOD score. We plan to develop the implementation of the ensemble distillation in the future.

Identifying the influential data among the low-contribution OOD samples in UL data can play a key role in improving the quality of photometric redshifts. Including influential data in training can alter the machine learning model significantly, reducing the fraction of incorrect estimation and/or the uncertainty of estimation (Charpiat et al. 2019; Pruthi et al. 2020). Since the influence of data relies on the model, the future method needs to have the model-dependent algorithm to assess the influence of data in a quantitative way. The highly influential data among the OOD samples requires labeling, i.e., acquiring spectroscopic redshifts, to be used as learning samples (Masters et al. 2015; Newman et al. 2015).

Understanding the nature of the low-contribution OOD samples will be also important in addition to the algorithm of evaluating the OOD score. A large fraction of the OOD samples might be produced by effects like source blending and observational artifacts. When selecting highly influential OOD samples, the new algorithm for the OOD score may need a step of filtering out the photometrically unreliable samples.

Semi-supervised learning methods (e.g., Ouali et al. 2020) might be an alternative approach to the supervised learning like the method presented here in the estimation of photometric redshifts. Semi-supervised learning models intrinsically do not have the issue of the OOD problem except for the case of physically OOD samples. However, the efficacy and reliability of the semi-supervised learning is a new challenge compared with supervised learning methods. Measuring the influence of data is still important because the performance of semi-supervised learning methods is strongly subject to the labeled samples and the weights of unlabeled samples.

Chapter 5

Summary and Discussion

The neural network is one of the representative machine learning architectures showing remarkable performances in a wide variety of fields and tasks. The astronomical applications of neural networks presented in this thesis show that 1) neural networks may achieve comparable or superior performances to the conventional methods for gravitational waveform generation and photometric redshift estimation, 2) the performance of neural network depends on the properties of training data, 3) neural network results can be overconfident and unreliable on samples drawn from OOD, and 4) open a new door for more reliable application of neural network to astrophysical problems.

In Chapter 2, we have built the DDS2S model generating merger-ringdown waveforms using the inspiral waveforms in a short time. We have examined the applicability of the waveforms by computing the overlap with EOBNR-based waveforms and performing the injection test. The accuracy of the DL-based waveforms is found to be better than 99.9% in most combinations of the masses, while a small number of outliers with overlap as small as 0.99 exists. In the injection test, we have recovered the event time of waveforms injected into real noise data with the conventional matched filtering engine of `PyCBC`. Regarding the speed of waveform generation, the DDS2S model has an advantage over other waveform approximants when computing a batch of multiple waveforms simultaneously. For computing a single waveform, EOB is faster than the DDS2S model, typically taking $O(10^{-2})$ seconds using a modern CPU core. However, the DDS2S model generates $\sim$ 1500 waveforms using pre-generated inspiral

waveforms in $O(1)$ seconds using NVIDIA GeForce GTX 1080, while EOB took $O(10)$ seconds. The disparity arises since the DL models are specialized for batch computations, which process multiple data at once.

The DDS2S model has been built to learn how to predict the output waveforms only from the given input waveforms without any specific physical information of the source binary system. Thus, we can readily extend this work to various systems of interest. For a more precise description of realistic physical binary systems, we need to have waveform models for more complex binaries: a wider range of the mass ratios, the spin of each component, eccentricity of the orbits. GWs from unbound orbit such as hyperbolic and parabolic encounters are also of great interest. Lastly, it is worthwhile to mention that recalibration of full IMR waveforms to increased amounts of NR waveform data is in progress in the community (The LSC-Virgo-KAGRA Observational Science Working Groups 2020). Our approach described in Chapter 2 can potentially be applied to more complex systems described above because DDS2S only depends on training data, not any assumptions or approximations on which other waveform models are based. Moreover, we have observed that $\sim$ 1000 training waveforms are sufficient for the model to reach the expected level of accuracy in Sec 2.4.3. Thus, as long as there is a sufficiently large number of training waveform samples for any systems or NR are given, DDS2S can be trained to generate accurate waveforms in principle.

In Chapter 3, we have improved the accuracy of neural networks for photometric redshift estimation and investigated how model performance varies regarding the distributions of samples in the input dimension spaces. Our investigation of the trained model's weakness guides us to issues that need to be addressed in order to improve the machine learning inference of photometric redshifts. First of all, more spectroscopic samples are required to improve the accuracy of the trained model. The accuracy of the machine learning inference might be degraded for the specific parts of mapping between the input and the redshift space due to the lack of enough training samples, which are generally considered close to the OOD samples (Beck et al. 2017). The mismatch between the training sample distribution and the test data distribution also results in the biased estimation of photometric redshifts (Rivera et al. 2018). Even when the test data incorporates only galaxies, the distribution of the input features in the test data

can become the out-of-distribution case as we examined in Section 3.5.1. Therefore, new machine learning models definitely need to include more spectroscopic samples as training data covering the large input and output (i.e., redshift) spaces.

In Chapter 4, we have presented a multi-stage training approach composed of supervised and unsupervised learnings to filter out OOD objects. Our model presented here successfully estimate photometric redshifts of galaxies even with the addition of the new network to our previous model presented in Chapter 3. As presented in Section 4.3.1, the proposed multi-step approach of this chapter reproduces a reliable estimation of photometric redshifts. The proposed method to detect the OOD data with respect to the photometric redshift inference model shows that the new model can clearly measure how much given samples deviate from the training samples as OOD objects when they are physically galaxy objects with the OOD properties (see Section 4.3.2). The majority of objects with the catastrophic redshift estimation is explained as data with a high OOD score (see Figure 4.5). The model estimating the OOD score cannot replace the model of classifying out galaxies from stars and quasars. As shown in Figure 4.10, a large fraction of stars and quasars show low OOD scores and high confidence values for photometric redshifts in our model. These objects have indistinguishable input features compared to the real galaxy objects in terms of the model perspective. Identifying the influential data among the OOD samples can play a key role in improving the quality of photometric redshifts. The influential data can alter the machine learning model significantly, reducing the fraction of incorrect estimation. Since the influence of data relies on the model, the future method needs to have a model-dependent algorithm to assess the influence of data in a quantitative way. The highly influential data requires labeling, i.e., acquiring spectroscopic redshifts, to be used as learning samples.

As we have shown throughout these sequential studies, neural network-based astronomical approaches may achieve similar or superior performances to the conventional methods. DDS2S model achieves comparable accuracies with EOBNR-based gravitational waveform templates as shown in Chapter 2, and MBRNN shows higher photometric accuracies than typical SED fitting based redshifts, as proved in Chapter 3. However, the stability and high performance of the models are restricted to the input dimension spaces extended by the training data. The

overlaps of the generated gravitational waveforms by DDS2S models plummet for the parameters drawn outside the training data. Besides, the results of the MBRNN for OOD objects are inaccurate and/or overconfident. The unintended test data not included in the training set spoil the credibility of the model.

The unstable performance of neural networks is a generally well-known problem in computer science. Hendricks and Gimpel (Hendrycks & Gimpel 2016) suggested a baseline model for OOD detection in the neural network. Lee et al (K. Lee et al. 2018) proposed a progressed method for detection of OOD examples using Mahalanobis distance assuming the trained network parameters can be fitted well by a class-conditional Gaussian distribution. Yu and Aizawa (Yu & Aizawa 2019) use unsupervised training with unlabeled samples as training data to endow the networks with the functionality of scoring and detecting OOD objects. Some of these example studies on OOD problems emphasize the awareness of the community for the untrustworthiness of the neural network and the movement towards more robust models.

In the field of Astronomy, however, this unreliability of neural networks regarding OOD has been disregarded shaded by the superior performance and the readily applicable characteristics of the models. Although numerous past studies have proved the advantages of using neural network-based approaches (Firth et al. 2003; Ball et al. 2008; Singal et al. 2011; Brescia et al. 2013; Laigle et al. 2017; Bilicki et al. 2018; Chong & Yang 2019), we haven't been able to find any astronomical research cases handling robustness or OOD problems concerning the neural network to the best of our knowledge. The data of interest for the astronomical application of neural networks should have unwanted samples since it is implausible to prepare training data covering entire input dimension spaces extended by real-world data. Hence, the superior performances of the neural network, which have been shown in the optimistic standpoint so far, are somewhat overestimated since well-controlled test samples are only used for the assessment of the model performance, filtering out OOD objects. Using this series of works as a springboard, we expect more stringent studies on the reliability of the neural network to be made for a practical application of the architecture in Astronomy.

References

Aasi, J., et al. (2013, Sep). Parameter estimation for compact binary coalescence signals with the first generation gravitational-wave detector network. *Phys. Rev. D*, *88*, 062001. Retrieved from https://link.aps.org/doi/10.1103/PhysRevD.88.062001 doi: 10.1103/PhysRevD.88.062001

Aasi, J., et al. (2015). Advanced LIGO. *Class. Quant. Grav.*, *32*, 074001. doi: 10.1088/0264-9381/32/7/074001

Abbott, B. P., et al. (2016). Observation of Gravitational Waves from a Binary Black Hole Merger. *Phys. Rev. Lett.*, *116*(6), 061102. doi: 10.1103/PhysRevLett.116.061102

Abbott, B. P., et al. (2017). The basic physics of the binary black hole merger GW150914. *Annalen Phys.*, *529*(1-2), 1600209. doi: 10.1002/andp.201600209

Abbott, R., et al. (2020, October). GWTC-2: Compact Binary Coalescences Observed by LIGO and Virgo During the First Half of the Third Observing Run. *arXiv e-prints*, arXiv:2010.14527.

Acernese, F., et al. (2015). Advanced Virgo: a second-generation interferometric gravitational wave detector. *Class. Quant. Grav.*, *32*(2), 024001. doi: 10.1088/0264-9381/32/2/024001

Aguado, D. S., Ahumada, R., Almeida, A., Anderson, S. F., Andrews, B. H., Anguiano, B., ... Zou, H. (2019, February). The Fifteenth Data Release of the Sloan Digital Sky Surveys: First Release of MaNGA-derived Quantities, Data Visualization Tools, and Stellar Library. *ApJS*, *240*(2), 23. doi: 10.3847/1538-4365/aaf651

Alam, S., Albareti, et al. (2015, July). The Eleventh and Twelfth Data Releases of the Sloan Digital Sky Survey: Final Data from SDSS-III. *ApJS*, *219*(1), 12. doi: 10.1088/0067-0049/219/1/12

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, *46*(3), 175-185. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/00031305.1992.10475879 doi: 10.1080/00031305.1992.10475879

Amon, A., Blake, C., Heymans, C., Leonard, C. D., Asgari, M., Bilicki, M., ... Wolf, C. (2018, 06). KiDS+2dFLenS+GAMA: testing the cosmological model with the EG statistic. *Monthly Notices of the Royal Astronomical Society*, *479*(3), 3422-3437. Retrieved from https://doi.org/10.1093/mnras/sty1624 doi: 10.1093/mnras/sty1624

Ball, N. M., Brunner, R. J., Myers, A. D., Strand, N. E., Alberts, S. L., & Tcheng, D. (2008). Robust machine learning applied to astronomical data sets. iii. probabilistic photometric redshifts for galaxies and quasars in the sdss and galex. *The Astrophysical Journal*, *683*(1), 12.

Banerji, M., Abdalla, F. B., Lahav, O., & Lin, H. (2008, 04). Photometric redshifts for the Dark Energy Survey and VISTA and implications for large-scale structure. *Monthly Notices of the Royal Astronomical Society*, *386*(3), 1219-1233. Retrieved from https://doi.org/10.1111/j.1365-2966.2008.13095.x doi: 10.1111/j.1365-2966.2008.13095.x

Barron, J. T. (2019, June). A general and adaptive robust loss function. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr).*

Beck, R., Lin, C. A., Ishida, E. E. O., Gieseke, F., de Souza, R. S., Costa-Duarte, M. V., ... Krone-Martins, A. (2017, July). On the realistic validation of photometric redshifts. *MNRAS*, *468*(4), 4323-4339. doi: 10.1093/mnras/stx687

Beck, R., Szapudi, I., Flewelling, H., Holmberg, C., Magnier, E., & Chambers, K. C. (2021, January). PS1-STRM: neural network source classification and photometric redshift catalogue for PS1 3π DR1. *MNRAS*, *500*(2), 1633-1644. doi: 10.1093/mnras/staa2587

Bengio, Y., Ducharme, R., & Vincent, P. (2001). A neural probabilistic language model. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems 13* (pp. 932–938). MIT Press. Retrieved from http://papers.nips.cc/paper/1839-a-neural-probabilistic-language-model.pdf

Berger, P., & Stein, G. (2019). A volumetric deep convolutional neural network for simulation of mock dark matter halo catalogues. *Monthly Notices of the Royal Astronomical Society*, *482*(3), 2861–2871.

Bilicki, M., Hoekstra, H., Brown, M., Amaro, V., Blake, C., Cavuoti, S., ... others (2018). Photometric redshifts for the kilo-degree survey-machine-learning analysis with artificial neural networks. *Astronomy & Astrophysics*, *616*, A69.

Blake, C., & Bridle, S. (2005, 11). Cosmology with photometric redshift surveys. *Monthly Notices of the Royal Astronomical Society*, *363*(4), 1329-1348. Retrieved from https://doi.org/10.1111/j.1365-2966.2005.09526.x doi: 10.1111/j.1365-2966.2005.09526.x

Blanchet, L., Damour, T., Esposito-Farese, G., & Iyer, B. R. (2004). Gravitational radiation from inspiralling compact binaries completed at the third post-Newtonian order. *Phys. Rev. Lett.*, *93*, 091101. doi: 10.1103/PhysRevLett.93.091101

Blanchet, L., Damour, T., Iyer, B. R., Will, C. M., & Wiseman, A. (1995). Gravitational radiation damping of compact binary systems to second postNewtonian order. *Phys. Rev. Lett.*, *74*, 3515-3518. doi: 10.1103/PhysRevLett.74.3515

Bohé, A., et al. (2016). An improved effective-one-body model of spinning, nonprecessing binary black holes for the era of gravitational-wave astrophysics with advanced detectors. *arXiv preprint arXiv:1611.03703*.

Bohé, A., Shao, L., Taracchini, A., Buonanno, A., Babak, S., Harry, I. W., ... Szilágyi, B. (2017, Feb). Improved effective-one-body model of spinning, nonprecessing binary black holes for the era of gravitational-wave astrophysics with advanced detectors. *Phys. Rev.*

D, *95*, 044028. Retrieved from https://link.aps.org/doi/10.1103/PhysRevD.95.044028 doi: 10.1103/PhysRevD.95.044028

Bolzonella, M., Miralles, J. M., & Pelló, R. (2000, November). Photometric redshifts based on standard SED fitting procedures. *A&A*, *363*, 476-492.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of compstat'2010* (p. 177-186). Retrieved from https://app.dimensions.ai/details/publication/pub.1017229575andhttp://leon.bottou.org/publications/pdf/compstat-2010.pdf doi: 10.1007/978-3-7908-2604-3_16

Boyle, M., et al. (2019). The SXS Collaboration catalog of binary black hole simulations.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. Retrieved from http://dx.doi.org/10.1023/A%3A1010933404324 doi: 10.1023/A:1010933404324

Brescia, M., Cavuoti, S., D'Abrusco, R., Longo, G., & Mercurio, A. (2013, jul). PHOTOMETRIC REDSHIFTS FOR QUASARS IN MULTI-BAND SURVEYS. *The Astrophysical Journal*, *772*(2), 140. Retrieved from https://doi.org/10.1088%2F0004-637x%2F772%2F2%2F140 doi: 10.1088/0004-637x/772/2/140

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020, 05). *Language models are few-shot learners.*

Brown, T. B., et al. (2020, May). Language Models are Few-Shot Learners. *arXiv e-prints*, arXiv:2005.14165.

Buonanno, A., & Damour, T. (2000). Transition from inspiral to plunge in binary black hole coalescences. *Phys. Rev.*, *D62*, 064015. doi: 10.1103/PhysRevD.62.064015

Cavuoti, S., Amaro, V., Brescia, M., Vellucci, C., Tortora, C., & Longo, G. (2017, February). METAPHOR: a machine-learning-based method for the probability density estimation of photometric redshifts. *MNRAS*, *465*(2), 1959-1973. doi: 10.1093/mnras/stw2930

Chambers, K. C., Magnier, E. A., Metcalfe, N., Flewelling, H. A., Huber, M. E., Waters, C. Z., ... Wyse, R. (2016, December). The Pan-STARRS1 Surveys. *arXiv e-prints*, arXiv:1612.05560.

Charpiat, G., Girard, N., Felardos, L., & Tarabalka, Y. (2019). Input similarity from the neural network perspective. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc. Retrieved from `https://proceedings.neurips.cc/paper/2019/file/c61f571dbd2fb949d3fe5ae1608dd48b-Paper.pdf`

Chen, B. H., Goto, T., Kim, S. J., Wang, T. W., Santos, D. J. D., Ho, S. C., ... others (2021). An active galactic nucleus recognition model based on deep neural network. *Monthly Notices of the Royal Astronomical Society*, *501*(3), 3951–3961.

Childress, M. J., Lidman, C., Davis, T. M., Tucker, B. E., Asorey, J., Yuan, F., ... Zhang, B. R. (2017, November). OzDES multifibre spectroscopy for the Dark Energy Survey: 3-yr results and first data release. *MNRAS*, *472*(1), 273-288. doi: 10.1093/mnras/stx1872

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Chong, K., De Wei, & Yang, A. (2019, September). Photometric Redshift Analysis using Supervised Learning Algorithms and Deep Learning. In *European physical journal web of conferences* (Vol. 206, p. 09006). doi: 10.1051/epjconf/201920609006

Chua, A. J. K., Galley, C. R., & Vallisneri, M. (2019). *Reduced-Order Modeling with Artificial Neurons for Gravitational-Wave Inference*. *Phys. Rev. Lett.*, *122*, 211101.

Colless, M., Dalton, G., Maddox, S., Sutherland, W., Norberg, P., Cole, S., ... Taylor, K. (2001, December). The 2dF Galaxy Redshift Survey: spectra and redshifts. *MNRAS*, *328*(4), 1039-1063. doi: 10.1046/j.1365-8711.2001.04902.x

Cool, R. J., Moustakas, J., Blanton, M. R., Burles, S. M., Coil, A. L., Eisenstein, D. J., ... Mendez, A. J. (2013, April). The PRIsm MUlti-object Survey (PRIMUS). II. Data Reduction and Redshift Fitting. *ApJ*, *767*(2), 118. doi: 10.1088/0004-637X/767/2/118

Cortes, C., & Vapnik, V. (1995). Support vector networks. *Machine Learning*, *20*, 273-297.

Csabai, I., Dobos, L., Trencséni, M., Herczegh, G., Józsa, P., Purger, N., ... Szalay, A. S. (2007, October). Multidimensional indexing tools for the virtual observatory. *Astronomische Nachrichten*, *328*(8), 852. doi: 10.1002/asna.200710817

Cui, X.-Q., Zhao, Y.-H., Chu, Y.-Q., Li, G.-P., Li, Q., Zhang, L.-P., ... Zou, S.-C. (2012, September). The Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST). *Research in Astronomy and Astrophysics*, *12*(9), 1197-1242. doi: 10.1088/1674-4527/12/9/003

Cutler, C., & Vallisneri, M. (2007, Nov). Lisa detections of massive black hole inspirals: Parameter extraction errors due to inaccurate template waveforms. *Phys. Rev. D*, *76*, 104018. Retrieved from https://link.aps.org/doi/10.1103/PhysRevD.76.104018 doi: 10.1103/PhysRevD.76.104018

D'Isanto, A., Cavuoti, S., Gieseke, F., & Polsterer, K. L. (2018, August). Return of the features. Efficient feature selection and interpretation for photometric redshifts. *A&A*, *616*, A97. doi: 10.1051/0004-6361/201833103

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., & Koltun, V. (2017a). CARLA: An open urban driving simulator. In *Proceedings of the 1st annual conference on robot learning* (pp. 1–16).

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., & Koltun, V. (2017b, 13–15 Nov). CARLA: An open urban driving simulator. In S. Levine, V. Vanhoucke, & K. Goldberg (Eds.), *Proceedings of the 1st annual conference on robot learning* (Vol. 78, pp. 1–16). PMLR. Retrieved from http://proceedings.mlr.press/v78/dosovitskiy17a.html

Droz, S., Knapp, D. J., Poisson, E., & Owen, B. J. (1999). Gravitational waves from inspiraling compact binaries: Validity of the stationary phase approximation to the Fourier transform. *Phys. Rev.*, *D59*, 124016. doi: 10.1103/PhysRevD.59.124016

Ellison, S. L., Teimoorinia, H., Rosario, D. J., & Mendel, J. T. (2016). The star formation rates of active galactic nuclei host galaxies. *Monthly Notices of the Royal Astronomical Society: Letters*, *458*(1), L34–L38.

Escamilla-Rivera, C., Quintero, M. A. C., & Capozziello, S. (2020). A deep learning approach to cosmological dark energy models. *Journal of Cosmology and Astroparticle Physics*, *2020*(03), 008.

Euclid Collaboration, Adam, R., Vannier, M., Maurogordato, S., Biviano, A., Adami, C., ... Zamorani, G. (2019, July). Euclid preparation. III. Galaxy cluster detection in the wide photometric survey, performance and algorithm selection. *A&A*, *627*, A23. doi: 10.1051/0004-6361/201935088

Favata, M. (2011, Jan). Conservative corrections to the innermost stable circular orbit (isco) of a kerr black hole: A new gauge-invariant post-newtonian isco condition, and the isco shift due to test-particle spin and the gravitational self-force. *Phys. Rev. D*, *83*, 024028. Retrieved from https://link.aps.org/doi/10.1103/PhysRevD.83.024028 doi: 10.1103/PhysRevD.83.024028

Firth, A. E., Lahav, O., & Somerville, R. S. (2003, 03). Estimating photometric redshifts with artificial neural networks. *Monthly Notices of the Royal Astronomical Society*, *339*(4), 1195-1202. Retrieved from https://doi.org/10.1046/j.1365-8711.2003.06271.x doi: 10.1046/j.1365-8711.2003.06271.x

Flanagan, É. É., & Hughes, S. A. (1998, April). Measuring gravitational waves from binary black hole coalescences. I. Signal to noise for inspiral, merger, and ringdown. *Phys. Rev. D*, *57*(8), 4535-4565. doi: 10.1103/PhysRevD.57.4535

Flewelling, H. A., Magnier, E. A., Chambers, K. C., Heasley, J. N., Holmberg, C., Huber, M. E., ... Shiao, B. (2020, November). The Pan-STARRS1 Database and Data Products. *ApJS*, *251*(1), 7. doi: 10.3847/1538-4365/abb82d

Fotopoulou, S., & Paltani, S. (2018, October). CPz: Classification-aided photometric-redshift estimation. *A&A*, *619*, A14. doi: 10.1051/0004-6361/201730763

Francisco Massa, R. M., & Aubry, M. (2016, September). Crafting a multi-task cnn for viewpoint estimation. In E. R. H. Richard C. Wilson & W. A. P. Smith (Eds.), *Proceedings of the british machine vision conference (bmvc)* (p. 91.1-91.12). BMVA Press. Retrieved from `https://dx.doi.org/10.5244/C.30.91` doi: 10.5244/C.30.91

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*, 193–202.

Galametz, A., Saglia, R., Paltani, S., Apostolakos, N., & Dubath, P. (2017, February). SED-dependent galactic extinction prescription for Euclid and future cosmological surveys. *A&A*, *598*, A20. doi: 10.1051/0004-6361/201629333

Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. In *Proceedings of the 34th international conference on machine learning-volume 70* (pp. 1243–1252).

Gers, F. A., Schmidhuber, J., & Cummins, F. (1999). Learning to forget: Continual prediction with lstm.

Ghosh, A., Urry, C. M., Wang, Z., Schawinski, K., Turp, D., & Powell, M. C. (2020). Galaxy morphology network: A convolutional neural network used to study morphology and quenching in 100,000 sdss and 20,000 candels galaxies. *The Astrophysical Journal*, *895*(2), 112.

Groné, A. (2017). *Hands-on machine learning with scikit-learn and tensorflow: Concepts, tools, and techniques to build intelligent systems* (1st ed.). O'Reilly Media, Inc.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017, 06–11 Aug). On calibration of modern neural networks. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning* (Vol. 70, pp. 1321–1330). PMLR. Retrieved from http://proceedings.mlr.press/v70/guo17a.html

Hasinger, G., Capak, P., Salvato, M., Barger, A. J., Cowie, L. L., Faisst, A., ... Yang, F. (2018, May). The DEIMOS 10K Spectroscopic Survey Catalog of the COSMOS Field. *ApJ*, *858*(2), 77. doi: 10.3847/1538-4357/aabacf

Healy, J., Lousto, C. O., Lange, J., O'Shaughnessy, R., Zlochower, Y., & Campanelli, M. (2019). The second rit binary black hole simulations catalog and its application to gravitational waves parameter estimation. *arXiv preprint arXiv:1901.02553*.

Healy, J., Lousto, C. O., Zlochower, Y., & Campanelli, M. (2017, oct). The RIT binary black hole simulations catalog. *Classical and Quantum Gravity*, *34*(22), 224001. Retrieved from https://doi.org/10.1088%2F1361-6382%2Faa91b1 doi: 10.1088/1361-6382/aa91b1

Hendrycks, D., & Gimpel, K. (2016). A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th international conference on learning representations, ICLR 2017, toulon, france, april 24-26, 2017, conference track proceedings*. OpenReview.net. Retrieved from https://openreview.net/forum?id=Hkg4TI9xl

Hendrycks, D., Mazeika, M., & Dietterich, T. G. (2019). Deep anomaly detection with outlier exposure. In *7th international conference on learning representations, ICLR 2019, new orleans, la, usa, may 6-9, 2019*. OpenReview.net. Retrieved from https://openreview.net/forum?id=HyxCxhRcY7

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*(8), 2554–2558. Retrieved from https://www.pnas.org/content/79/8/2554 doi: 10.1073/pnas.79.8.2554

Ioffe, S., & Szegedy, C. (2015, 07–09 Jul). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In F. Bach & D. Blei (Eds.), *Proceedings of the 32nd international conference on machine learning* (Vol. 37, pp. 448–456). Lille, France: PMLR. Retrieved from http://proceedings.mlr.press/v37/ioffe15.html

Ivezić, Ž., Kahn, S. M., Tyson, J. A., Abel, B., Acosta, E., Allsman, R., ... Zhan, H. (2019, March). LSST: From Science Drivers to Reference Design and Anticipated Data Products. *ApJ*, *873*(2), 111. doi: 10.3847/1538-4357/ab042c

Jani, K., Healy, J., Clark, J. A., London, L., Laguna, P., & Shoemaker, D. (2016, sep). Georgia tech catalog of gravitational waveforms. *Classical and Quantum Gravity*, *33*(20), 204001. Retrieved from https://doi.org/10.1088%2F0264-9381%2F33%2F20%2F204001 doi: 10.1088/0264-9381/33/20/204001

Jones, D. H., Read, M. A., Saunders, W., Colless, M., Jarrett, T., Parker, Q. A., ... Williams, M. (2009, October). The 6dF Galaxy Survey: final redshift release (DR3) and southern large-scale structures. *MNRAS*, *399*(2), 683-698. doi: 10.1111/j.1365-2966.2009.15338.x

Jones, D. M., & Heavens, A. F. (2019, February). Bayesian photometric redshifts of blended sources. *MNRAS*, *483*(2), 2487-2505. doi: 10.1093/mnras/sty3279

Kaiser, N., Burgett, W., Chambers, K., Denneau, L., Heasley, J., Jedicke, R., ... Tonry, J. (2010). The Pan-STARRS wide-field optical/NIR imaging survey. In *Proc. SPIE* (Vol. 7733, p. 77330E). doi: 10.1117/12.859188

Kalchbrenner, N., & Blunsom, P. (2013, October). Recurrent continuous translation models. Seattle: Association for Computational Linguistics.

Keller, S. C., Schmidt, B. P., Bessell, M. S., Conroy, P. G., Francis, P., Granlund, A., ... Waterson, M. F. (2007, May). The SkyMapper Telescope and The Southern Sky Survey. *PASA*, *24*(1), 1-12. doi: 10.1071/AS07001

Korytov, D., Hearin, A., Kovacs, E., Larsen, P., Rangel, E., Hollowed, J., ... (The LSST Dark Energy Science Collaboration (2019, December). CosmoDC2: A Synthetic Sky Catalog for Dark Energy Science with LSST. *ApJS*, *245*(2), 26. doi: 10.3847/1538-4365/ab510c

Krizhevsky, A. (2012, 05). Learning multiple layers of features from tiny images. *University of Toronto*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 25, pp. 1097–1105). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017, May). Imagenet classification with deep convolutional neural networks. *Commun. ACM*, *60*(6), 84–90. Retrieved from https://doi.org/10.1145/3065386 doi: 10.1145/3065386

Kumar, P., Chu, T., Fong, H., Pfeiffer, H. P., Boyle, M., Hemberger, D. A., ... Szilagyi, B. (2016, May). Accuracy of binary black hole waveform models for aligned-spin binaries. *Phys. Rev. D*, *93*, 104050. Retrieved from https://link.aps.org/doi/10.1103/PhysRevD.93.104050 doi: 10.1103/PhysRevD.93.104050

Kundu, A., Li, Y., & Rehg, J. M. (2018). 3d-rcnn: Instance-level 3d object reconstruction via render-and-compare. In *Cvpr*.

Laigle, C., Pichon, C., Arnouts, S., McCracken, H. J., Dubois, Y., Devriendt, J., ... Vibert, D. (2017, 11). COSMOS2015 photometric redshifts probe the impact of filaments on galaxy properties. *Monthly Notices of the Royal Astronomical Society*, *474*(4), 5437-5458. Retrieved from https://doi.org/10.1093/mnras/stx3055 doi: 10.1093/mnras/stx3055

Le Fèvre, O., Cassata, P., Cucciati, O., Garilli, B., Ilbert, O., Le Brun, V., ... Zucca, E. (2013, November). The VIMOS VLT Deep Survey final data release: a spectroscopic sample of 35 016 galaxies and AGN out to z ~6.7 selected with $17.5 \leq i_{AB} \leq 24.75$. *A&A*, *559*, A14. doi: 10.1051/0004-6361/201322179

LeCun, Y., & Cortes, C. (2010). MNIST handwritten digit database. http://yann.lecun.com/exdb/mnist/. Retrieved 2016-01-14 14:24:11, from `http://yann.lecun.com/exdb/mnist/`

Lee, J., Oh, S. H., Kim, K., Cho, G., Oh, J. J., Son, E. J., & Lee, H. M. (2021, Jun). Deep learning model on gravitational waveforms in merging and ringdown phases of binary black hole coalescences. *Phys. Rev. D*, *103*, 123023. Retrieved from `https://link.aps.org/doi/10.1103/PhysRevD.103.123023` doi: 10.1103/PhysRevD.103.123023

Lee, K., Lee, K., Lee, H., & Shin, J. (2018). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Proceedings of the 32nd international conference on neural information processing systems* (p. 7167–7177). Red Hook, NY, USA: Curran Associates Inc.

Levinson, J., Askeland, J., Becker, J., Dolson, J., Held, D., Kammel, S., ... others (2011). Towards fully autonomous driving: Systems and algorithms. In *2011 ieee intelligent vehicles symposium (iv)* (pp. 163–168).

Liang, S., Li, Y., & Srikant, R. (2017). Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*.

Lilly, S. J., Le Brun, V., Maier, C., Mainieri, V., Mignoli, M., Scodeggio, M., ... Taniguchi, Y. (2009, October). The zCOSMOS 10k-Bright Spectroscopic Sample. *ApJS*, *184*(2), 218-229. doi: 10.1088/0067-0049/184/2/218

Lilly, S. J., Le Fèvre, O., Renzini, A., Zamorani, G., Scodeggio, M., Contini, T., ... Zucca, E. (2007, September). zCOSMOS: A Large VLT/VIMOS Redshift Survey Covering 0 ¡ z ¡ 3 in the COSMOS Field. *ApJS*, *172*(1), 70-85. doi: 10.1086/516589

Liu, P. L. (2019). *Multimodal regression — beyond l1 and l2 loss.* Retrieved 2019-09-30, from https://towardsdatascience.com/anchors-and-multi-bin-loss-for-multi-modal-target-regression-647ea1974617

Lovelace, G., et al. (2016). Modeling the source of GW150914 with targeted numerical-relativity simulations. *Class. Quant. Grav.*, *33*(24), 244002. doi: 10.1088/0264-9381/33/24/244002

LSST Dark Energy Science Collaboration (LSST DESC), Abolfathi, B., Alonso, D., Armstrong, R., Aubourg, É., Awan, H., ... Zuntz, J. (2021, March). The LSST DESC DC2 Simulated Sky Survey. *ApJS*, *253*(1), 31. doi: 10.3847/1538-4365/abd62c

Luong, M.-T., Le, Q. V., Sutskever, I., Vinyals, O., & Kaiser, L. (2015). Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*.

Malinin, A., Mlodozeniec, B., & Gales, M. (2020). Ensemble distribution distillation. In *International conference on learning representations.* Retrieved from https://openreview.net/forum?id=BygSP6Vtvr

Masters, D., Capak, P., Stern, D., Ilbert, O., Salvato, M., Schmidt, S., ... Cavuoti, S. (2015, 10). Mapping the galaxy color-redshift relation: Optimal photometric redshift calibration strategies for cosmology surveys. *The Astrophysical Journal*, *813*. doi: 10.1088/0004-637X/813/1/53

Masters, D. C., Stern, D. K., Cohen, J. G., Capak, P. L., Rhodes, J. D., Castander, F. J., & Paltani, S. (2017, June). The Complete Calibration of the Color-Redshift Relation (C3R2) Survey: Survey Overview and Data Release 1. *ApJ*, *841*(2), 111. doi: 10.3847/1538-4357/aa6f08

Masters, D. C., Stern, D. K., Cohen, J. G., Capak, P. L., Stanford, S. A., Hernitschek, N., ... Fotopoulou, S. (2019, June). The Complete Calibration of the Color-Redshift Relation (C3R2) Survey: Analysis and Data Release 2. *ApJ*, *877*(2), 81. doi: 10.3847/1538-4357/ab184d

Mousavian, A., Anguelov, D., Flynn, J., & Kosecka, J. (2017, July). 3d bounding box estimation using deep learning and geometry. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr)*.

Nallapati, R., Zhou, B., Gulcehre, C., Xiang, B., et al. (2016). Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Newman, J. A., Abate, A., Abdalla, F. B., Allam, S., Allen, S. W., Ansari, R., ... Zentner, A. R. (2015). Spectroscopic needs for imaging dark energy experiments. *Astroparticle Physics*, *63*, 81-100. Retrieved from https://www.sciencedirect.com/science/article/pii/S0927650514000917 (Dark Energy and CMB) doi: https://doi.org/10.1016/j.astropartphys.2014.06.007

Newman, J. A., Cooper, M. C., Davis, M., Faber, S. M., Coil, A. L., Guhathakurta, P., ... Schiavon, R. P. (2013, September). The DEEP2 Galaxy Redshift Survey: Design, Observations, Data Reduction, and Redshifts. *ApJS*, *208*(1), 5. doi: 10.1088/0067-0049/208/1/5

Nishizawa, A. J., Hsieh, B.-C., Tanaka, M., & Takata, T. (2020, February). Photometric Redshifts for the Hyper Suprime-Cam Subaru Strategic Program Data Release 2. *arXiv e-prints*, arXiv:2003.01511.

Nitz, A., et al. (2018, November). *gwastro/pycbc: Pycbc v1.13.1 release.* Retrieved from https://doi.org/10.5281/zenodo.1490104 doi: 10.5281/zenodo.1490104

Ouali, Y., Hudelot, C., & Tami, M. (2020). An overview of deep semi-supervised learning. *ArXiv*, *abs/2006.05278*.

Pasquet, J., Bertin, E., Treyer, M., Arnouts, S., & Fouchez, D. (2019, January). Photometric redshifts from SDSS images using a convolutional neural network. *A&A*, *621*, A26. doi: 10.1051/0004-6361/201833617

Pasquet-Itam, J., & Pasquet, J. (2018, April). Deep learning approach for classifying, detecting and predicting photometric redshifts of quasars in the Sloan Digital Sky Survey stripe 82. *A&A*, *611*, A97. doi: 10.1051/0004-6361/201731106

Perraudin, N., Defferrard, M., Kacprzak, T., & Sgier, R. (2019). Deepsphere: Efficient spherical convolutional neural network with healpix sampling for cosmological applications. *Astronomy and Computing*, *27*, 130–146.

Piscopo, M. L., Spannowsky, M., & Waite, P. (2019). Solving differential equations with neural networks: Applications to the calculation of cosmological phase transitions. *Physical Review D*, *100*(1), 016002.

Planck Collaboration, Abergel, A., Ade, P. A. R., Aghanim, N., Alves, M. I. R., Aniano, G., ... Zonca, A. (2014, November). Planck 2013 results. XI. All-sky model of thermal dust emission. *A&A*, *571*, A11. doi: 10.1051/0004-6361/201323195

Pruthi, G., Liu, F., Kale, S., & Sundararajan, M. (2020). Estimating training data influence by tracing gradient descent. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 19920–19930). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2020/file/e6385d39ec9394f2f3a354d9d2b88eec-Paper.pdf

Pürrer, M. (2014). Frequency domain reduced order models for gravitational waves from aligned-spin compact binaries. *Class. Quant. Grav.*, *31*(19), 195010. doi: 10.1088/0264-9381/31/19/195010

Pürrer, M. (2016). Frequency domain reduced order model of aligned-spin effective-one-body waveforms with generic mass-ratios and spins. *Phys. Rev.*, *D93*(6), 064041. doi: 10.1103/PhysRevD.93.064041

Quinlan, J. R. (1986, March). Induction of decision trees. *Mach. Learn.*, *1*(1), 81–106. Retrieved from https://doi.org/10.1023/A:1022643204877 doi: 10.1023/A:1022643204877

Rafelski, M., Teplitz, H. I., Gardner, J. P., Coe, D., Bond, N. A., Koekemoer, A. M., ... Voyer, E. N. (2015, July). UVUDF: Ultraviolet Through Near-infrared Catalog and Photometric Redshifts of Galaxies in the Hubble Ultra Deep Field. *AJ*, *150*(1), 31. doi: 10.1088/0004-6256/150/1/31

Reiman, D. (2020). *Neural probabilistic modeling for astrophysics and galaxy evolution* (Unpublished doctoral dissertation). UC Santa Cruz.

Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., Depristo, M., ... Lakshminarayanan, B. (2019). Likelihood ratios for out-of-distribution detection. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 14707–14718). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/9611-likelihood-ratios-for-out-of-distribution-detection.pdf

Rines, K., Geller, M. J., Diaferio, A., & Kurtz, M. J. (2013, April). Measuring the Ultimate Halo Mass of Galaxy Clusters: Redshifts and Mass Profiles from the Hectospec Cluster Survey (HeCS). *ApJ*, *767*(1), 15. doi: 10.1088/0004-637X/767/1/15

Rivera, J. D., Moraes, B., Merson, A. I., Jouvel, S., Abdalla, F. B., & Abdalla, M. C. B. (2018, July). Degradation analysis in the estimation of photometric redshifts from non-representative training sets. *MNRAS*, *477*(4), 4330-4347. doi: 10.1093/mnras/sty880

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65–386.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1: Foundations* (p. 318–362). Cambridge, MA, USA: MIT Press.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, *115*(3), 211-252. doi: 10.1007/s11263-015-0816-y

Ryou, S., Jeong, S., & Perona, P. (2019, nov). Anchor loss: Modulating loss scale based on prediction difficulty. In *2019 ieee/cvf international conference on computer vision (iccv)* (p. 5991-6000). Los Alamitos, CA, USA: IEEE Computer Society. Retrieved from https://doi.ieeecomputersociety.org/10.1109/ICCV.2019.00609 doi: 10.1109/ICCV.2019.00609

Salvato, M., Ilbert, O., & Hoyle, B. (2019, June). The many flavours of photometric redshifts. *Nature Astronomy*, *3*, 212-222. doi: 10.1038/s41550-018-0478-0

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(3), 210-229. doi: 10.1147/rd.33.0210

Sanchez, C., Carrasco Kind, M., Lin, H., Miquel, R., Abdalla, F. B., Amara, A., ... Zuntz, J. (2014, 10). Photometric redshift analysis in the Dark Energy Survey Science Verification data. *Monthly Notices of the Royal Astronomical Society*, *445*(2), 1482-1506. Retrieved from `https://doi.org/10.1093/mnras/stu1836` doi: 10.1093/mnras/stu1836

Schindler, J.-T., Fan, X., Huang, Y.-H., Yue, M., Yang, J., Hall, P. B., ... Rees, J. M. (2019, July). The Extremely Luminous Quasar Survey in the Pan-STARRS 1 Footprint (PS-ELQS). *ApJS*, *243*(1), 5. doi: 10.3847/1538-4365/ab20d0

Scodeggio, M., Guzzo, L., Garilli, B., Granett, B. R., Bolzonella, M., de la Torre, S., ... Percival, W. J. (2018, January). The VIMOS Public Extragalactic Redshift Survey (VIPERS). Full spectroscopic data and auxiliary information release (PDR-2). *A&A*, *609*, A84. doi: 10.1051/0004-6361/201630114

Senior, A., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., ... Hassabis, D. (2020, 01). Improved protein structure prediction using potentials from deep learning. *Nature*, *577*, 1-5. doi: 10.1038/s41586-019-1923-7

Serrà, J., Álvarez, D., Gómez, V., Slizovskaia, O., Núñez, J. F., & Luque, J. (2020). Input complexity and out-of-distribution detection with likelihood-based generative models. In *International conference on learning representations.* Retrieved from `https://openreview.net/forum?id=SyxIWpVYvr`

Singal, J., Shmakova, M., Gerke, B., Griffith, R. L., & Lotz, J. (2011, May). The Efficacy of Galaxy Shape Parameters in Photometric Redshift Estimation: A Neural Network Approach. *PASP*, *123*(903), 615. doi: 10.1086/660155

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In *Proceedings of the 25th international conference on neural information processing systems - volume 2* (p. 2951–2959). Red Hook, NY, USA: Curran Associates Inc.

Sola, J., & Sevilla, J. (1997). Importance of input data normalization for the application of neural networks to complex industrial problems. *IEEE Transactions on Nuclear Science*, *44*(3), 1464-1468.

Sturani, R., Fischetti, S., Cadonati, L., Guidi, G. M., Healy, J., Shoemaker, D., & Viceré, A. (2010, aug). Complete phenomenological gravitational waveforms from spinning coalescing binaries. *Journal of Physics: Conference Series*, *243*, 012007. Retrieved from `https://doi.org/10.1088%2F1742-6596%2F243%2F1%2F012007` doi: 10.1088/1742-6596/243/1/012007

Su, H., Qi, C. R., Li, Y., & Guibas, L. J. (2015). Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *2015 ieee international conference on computer vision (iccv)* (p. 2686-2694). doi: 10.1109/ICCV.2015.308

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proceedings of the 27th international conference on neural information processing systems - volume 2* (pp. 3104–3112). Cambridge, MA, USA: MIT Press. Retrieved from `http://dl.acm.org/citation.cfm?id=2969033.2969173`

Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015). Going deeper with convolutions. In *2015 ieee conference on computer vision and pattern recognition (cvpr)* (p. 1-9). doi: 10.1109/CVPR.2015.7298594

Tachibana, Y., & Miller, A. A. (2018, December). A Morphological Classification Model to Identify Unresolved PanSTARRS1 Sources: Application in the ZTF Real-time Pipeline. *PASP*, *130*(994), 128001. doi: 10.1088/1538-3873/aae3d9

Tachibana, Y., & Miller, A. A. (2018, nov). A morphological classification model to identify unresolved PanSTARRS1 sources: Application in the ZTF real-time pipeline. *Publications*

of the Astronomical Society of the Pacific, *130*(994), 128001. Retrieved from https://doi.org/10.1088/1538-3873/aae3d9 doi: 10.1088/1538-3873/aae3d9

Tanaka, M. (2015, March). Photometric Redshift with Bayesian Priors on Physical Properties of Galaxies. *ApJ*, *801*(1), 20. doi: 10.1088/0004-637X/801/1/20

Tanaka, M., Coupon, J., Hsieh, B.-C., Mineo, S., Nishizawa, A. J., Speagle, J., ... Murayama, H. (2018, January). Photometric redshifts for Hyper Suprime-Cam Subaru Strategic Program Data Release 1. *PASJ*, *70*, S9. doi: 10.1093/pasj/psx077

Taracchini, A., et al. (2014). Effective-one-body model for black-hole binaries with generic mass ratios and spins. *Phys. Rev.*, *D89*(6), 061502. doi: 10.1103/PhysRevD.89.061502

The LSC-Virgo-KAGRA Observational Science Working Groups. (2020). *The lsc-virgo-kagra observational science white paper* (Tech. Rep. No. T2000424).

Tonry, J. L., Stubbs, C. W., Lykke, K. R., Doherty, P., Shivvers, I. S., Burgett, W. S., ... Wainscoat, R. J. (2012, May). The Pan-STARRS1 Photometric System. *ApJ*, *750*(2), 99. doi: 10.1088/0004-637X/750/2/99

Torrisi, M., Pollastri, G., & Le, Q. (2020). Deep learning methods in protein structure prediction. *Computational and Structural Biotechnology Journal*, *18*, 1301 - 1310. Retrieved from http://www.sciencedirect.com/science/article/pii/S2001037019304441 doi: https://doi.org/10.1016/j.csbj.2019.12.011

Tredcr, I. J. (1975). C.w. misner, k.s. thorne, j.a. wheeler: Gravitation. w.h. freeman and company limited, reading (england) 1973 â xxvi + 1279 seiten, preis Â£ 19.20 (clothbound); Â£ 8.60 (paperbound). *Astronomische Nachrichten*, *296*(1), 45-46. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/asna.19752960110 doi: 10.1002/asna.19752960110

Trump, J. R., Impey, C. D., Elvis, M., McCarthy, P. J., Huchra, J. P., Brusa, M., ... Smolčić, V. (2009, May). The COSMOS Active Galactic Nucleus Spectroscopic Survey. I. XMM-Newton Counterparts. *ApJ*, *696*(2), 1195-1212. doi: 10.1088/0004-637X/696/2/1195

Turin, G. (1960, June). An introduction to matched filters. *IRE Transactions on Information Theory*, *6*(3), 311-329. doi: 10.1109/TIT.1960.1057571

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460. Retrieved from http://www.jstor.org/stable/2251299

Tyson, J. A. (2002, December). Large Synoptic Survey Telescope: Overview. In J. A. Tyson & S. Wolff (Eds.), *Survey and other telescope technologies and discoveries* (Vol. 4836, p. 10-20). doi: 10.1117/12.456772

Urrutia, T., Wisotzki, L., Kerutt, J., Schmidt, K. B., Herenz, E. C., Klar, J., ... Weilbacher, P. M. (2019, April). The MUSE-Wide Survey: survey description and first data release. *A&A*, *624*, A141. doi: 10.1051/0004-6361/201834656

Usman, S. A., et al. (2016). *The PyCBC search for gravitational waves from compact binary coalescence*. *Class. Quant. Grav.*, *33*(21), 215004. doi: 10.1088/0264-9381/33/21/215004

Van Der Sluys, M., Raymond, V., Mandel, I., Röver, C., Christensen, N., Kalogera, V., ... Vecchio, A. (2008). Parameter estimation of spinning binary inspirals using markov chain monte carlo. *Classical and Quantum Gravity*, *25*(18), 184011.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In I. Guyon et al. (Eds.), *Advances in neural information processing systems* (Vol. 30, pp. 5998–6008). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017, June). Attention Is All You Need. *arXiv e-prints*, arXiv:1706.03762.

Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., & Saenko, K. (2015). Sequence to sequence-video to text. In *Proceedings of the ieee international conference on computer vision* (pp. 4534–4542).

Walcher, J., Groves, B., Budavári, T., & Dale, D. (2011, January). Fitting the integrated spectral energy distributions of galaxies. *Ap&SS*, *331*, 1-52. doi: 10.1007/s10509-010-0458-z

Wei, W., & Huerta, E. A. (2020, January). Gravitational wave denoising of binary black hole mergers with deep learning. *Physics Letters B*, *800*, 135081. doi: 10.1016/j.physletb.2019.135081

Williams, D., Siong Heng, I., Gair, J., A Clark, J., & Khamesra, B. (2019, 03). *A precessing numerical relativity waveform surrogate model for binary black holes: A gaussian process regression approach.*

Yu, Q., & Aizawa, K. (2019, October). Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *Proceedings of the ieee/cvf international conference on computer vision (iccv).*

Yu-yan, J. (2010). Selective ensemble learning algorithm. In *2010 international conference on electrical and control engineering* (p. 1859-1862). doi: 10.1109/iCECE.2010.457

Zhang, Y., Ma, H., Peng, N., Zhao, Y., & Wu, X.-b. (2013, August). Estimating Photometric Redshifts of Quasars via the k-nearest Neighbor Approach Based on Large Survey Databases. *AJ*, *146*(2), 22. doi: 10.1088/0004-6256/146/2/22

Zhang, Y., & Zhao, Y. (2015, May). Astronomy in the Big Data Era. *Data Science Journal*, *14*, 11. doi: 10.5334/dsj-2015-011

Zheng, H., Yang, Z., Liu, W., Liang, J., & Li, Y. (2015). Improving deep neural networks using softplus units. In *2015 international joint conference on neural networks (ijcnn)* (p. 1-4). doi: 10.1109/IJCNN.2015.7280459

Zhou, Z.-H. (2009). Ensemble learning. In S. Z. Li & A. Jain (Eds.), *Encyclopedia of biometrics* (pp. 270–273). Boston, MA: Springer US. Retrieved from `https://doi.org/10.1007/978-0-387-73003-5_293` doi: 10.1007/978-0-387-73003-5_293

요 약

인공신경망(neural network)은 많은 분야에서 높은 성능을 보이고있는 대표적인 기계학습(machine learning) 알고리즘의 하나이다. 우리는 본 논문을 통해 1) 인경신경망을 중력파 파형 생성, 측광 적색편이 추정 분야에 적용해 전통적인 연구 방법들과 동일하거나 더 높은 성능을 낼 수 있음을 보이며, 2) 인공신경망의 성능이 데이터의 어떤 성질에 따라 편차를 보이는지 연구하고, 3) 훈련데이터와는 다른 종류의 데이터가 인공신경망의 성능에 어떠한 영향을 주는지 밝히고, 4) 천문학 분야에 대한 인공신경망의 보다 신뢰도 높은 연구의 문을 연다. 정확하고 많은 양의 중력파 파형은 작은 세기를 갖는 중력파를 관측하기 위한 필수 요소 중 하나로, 수치상대론(numerical relativity)을 이용해 정확한 계산이 가능하다. 하지만 수치상대론은 너무 많은 계산양을 요해 많은 양의 파형을 빠른 시간안에 계산할 수 없다. 우리는 회귀적 인공신경망(recurrent neural network)을 응용해 새로운 구조의 모델을 설계하고, 해당 모델을 이용해 $O(1)$초 시간 안에 99% 이상의 정확도로 대략 1500개의 파형을 생성해 낸다. 기계학습 파형의 실천적 유효성을 검증하기 위한 방법으로 모수추정이 사용되었다. 모수추정은 실제 레이저 간섭계 중력파 관측소(Laser Interferometer Gravitational-Wave Observatory)의 잡음 데이터에 삽입된 모의 중력파 신호의 모수를 기계학습을 통해 생성한 파형들을 이용해 90%의 신뢰구간 내의 범위에서 추정하는 방식으로 구성된다. 이에 더해, 우리는 은하의 측광 적색편이(photometric redshift)를 추정하는 인공신경망 모델을 설계해 인공신경망의 적용이 단순히 하나의 천문분야에만 국한되지 않음을 증명한다. 해당 모델은 제한된 데이터인 은하의 색(color)과 관련된 데이터만을 이용해 아주 높은 정확도로 은하의 분광 적색편이(spectroscopic redshift)를 근사한다. 동시에, 우리는 인공신경망이 입력차원공간(input dimension space)에서 작은 데이터 밀도를 갖는 곳에 위치하는 데이터들과 훈련에 사용되지 않은 준항성체(quasar)와 별(star)에 속하는 데이터에 대해 낮은 정확도를 가짐을 보인다. 천문분야에서 인공신경망의 현실적인 적용을 위해 우리는 해당 종류의 데이터를 분포 외 데이터(out-of-distribution)로 정의하고, 분포 외 데이터가 모델의 입력으로 주어졌을 때, 우리는 모델이 기존 업무 성능을 유지하며 동시에 해당 데이터를 98%의 높은 정확도로 검출할 수 있는 기능을 비지도학습을 이용해 구현한다.

주요어: 기계학습: 인공신경망: 중력파: 은하: 준항성체: 별: 분포 외 데이터

학 번: 2014-22385