Estimation of Photometric Redshifts. II. Identification of Out-of-distribution Data with Neural Networks

Joongoo Lee1,2 and Min-Su Shin1

1 Korea Astronomy and Space Science Institute, 776, Daedeokdae-ro, Yuseong-gu, Daejeon 34055, Republic of Korea; 2 Department of Physics and Astronomy, Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea

Abstract

In this study, we propose a three-stage training approach of neural networks for both photometric redshift estimation of galaxies and detection of out-of-distribution (OOD) objects. Our approach comprises supervised and unsupervised learning, which enables using unlabeled (UL) data for OOD detection in training the networks. Employing the UL data, which is the data set 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 (LOOD) samples such as stars and quasars. Our model successfully produces photometric redshifts matched with spectroscopic redshifts for the ID samples and identifies well the LOOD objects with more than 98% accuracy. Although quantitative assessment with the UL samples is impracticable owing to the lack of labels and spectroscopic redshifts, we also find that our model successfully estimates reasonable photometric redshifts for ID-like UL samples and filter OOD-like UL objects.3

1. Introduction

Future astronomical surveys, such as the Legacy Survey of Space and Time (Ivezić et al. 2019), Euclid (Laureijs et al. 2010), and Nancy Grace Roman Space Telescope (Spergel et al. 2013), are expected to explore billions of galaxies whose redshifts should be reliably estimated. The redshifts of galaxies are prerequisites in most extragalactic and cosmological studies (Blake & Bridle 2005; Bordoloi et al. 2010; Masters et al. 2015). Although spectroscopic measurements yield accurate redshifts of distant galaxies with negligible error (Beck et al. 2016), obtaining spectroscopic redshifts is excessively time-consuming. Moreover, the number of objects to which spectroscopy is applicable is limited because high-quality spectroscopic data are expensive to acquire, particularly for faint objects. The cost of spectroscopic redshifts leads to the use of photometric redshifts as alternatives. Photometric redshifts are less time intensive and more widely applicable than spectroscopic ones, although their accuracy is generally worse than that of spectroscopic redshifts (Salvato et al. 2019).

Template-based spectral energy distribution (SED) fitting (Bolzonella et al. 2000) and machine-learning inference are two representative approaches for estimating photometric redshifts (Cavuoti et al. 2017). Although the template-based SED fitting provides extensive coverage of redshifts and photometric properties, it can be more computationally expensive than machine-learning inference owing to the nature of brute-force search. Besides, the reliance on prior knowledge on SEDs may induce bias in results (Walcher et al. 2011; Tanaka 2015). Meanwhile, machine-learning approaches enable quick inference of photometric redshifts after training phases (Salvato et al. 2019) as proven in many past studies (Carlilès et al. 2010; Gerdes et al. 2010; D’Isanto & Polsterer 2018; Pasquet-Itam & Pasquet 2018). Despite this advantage of the machine-learning approaches, the reliability and performance of machine-learning models are guaranteed only when test data follow the distribution of training data in both input and target (i.e., redshift) spaces as in-distribution (ID) data (Goodfellow et al. 2015; Amodei et al. 2016).

It is practically impossible to prepare a training set with complete coverage of the entire input and target (i.e., redshift) spaces extended by the expected real-world data (Domingos 2012; Yuille & Liu 2021). The underrepresented or unseen galaxies residing outside the property range of training samples for inference of photometric redshifts are drawn from out-of-distribution (OOD) data with respect to given machine-learning models (Hendrycks & Gimpel 2017; Liang et al. 2018; Lee et al. 2018a, 2018b; Ren et al. 2019). The OOD objects in the in-distribution of photometric redshifts can be categorized into two types. One type corresponds to objects such as quasars (QSOs) and stars physically different from galaxies. We call them physically OOD objects. Although the distribution of the physically OOD samples generally deviates from that of ID samples in the input space, some may exist where the ID samples densely populate. For example, the colors of QSOs at some redshifts can be similar to those of galaxies at different redshifts. Hence, the physically OOD samples may be viewed as ID samples from the perspective of the trained model. The second type comprises galaxies unseen or underrepresented in training data (hereinafter underrepresented galaxies). The photometric redshifts of underrepresented galaxies cannot be successfully inferred by a model owing to the nature of inductive reasoning in machine learning.

In our previous study (Lee & Shin 2021, hereafter Paper I) we focused on the accurate photometric redshift estimation of ID samples that are galaxies well represented by training data, and the trained model successfully inferred photometric redshifts with ID test samples. However, we found that both nongalaxy objects and underrepresented galaxies could have an overconfident estimation of photometric redshifts, although the estimation was incorrect. Confidence is a measure of how
confident a neural network is about its inference, and it is defined as the maximum value of the probability output for a given sample. Silent failure of trained networks by estimating incorrect photometric redshifts with high confidence for the OOD objects proves that the typical confidence is an unreliable indicator for detecting OODs and may lead to wrong interpretation (Nguyen et al. 2015). Owing to these unreliable results and the vulnerability of trained networks, alternative approaches for OOD detection are required.

In this second paper of a series, we propose a multistage-training approach to handle dual tasks: photometric redshift estimation and OOD scoring/detection. The proposed approach employs unsupervised learning for OOD scoring/detection introduced in Yu & Aizawa (2019), in addition to supervised learning for redshift estimation described in Paper I. Supervised learning requiring labeled samples (i.e., training samples of galaxies with spectroscopic redshifts) in training the model shows a remarkable performance on ID data (Caruana & Niculescu-Mizil 2006; see Paper I). We employ unsupervised learning, which trains models with a loss function computable without sample labels (Yamanishi & Takeuchi 2001; Chui et al. 2004) for using unlabeled (UL) data, presumably containing both ID and OOD data. For inference, we use networks’ outputs from different timelines of the training process to evaluate photometric redshifts and OOD scores. Since the networks trained as such can both detect OOD objects and estimate photometric redshifts, we expect a trained model to perform better with real-world data than a model without handling the OOD samples.

We also expect that our approach employing the UL training data for OOD object detection can achieve a more robust performance regarding the OOD data by handling the more diverse possibility of OOD properties with the UL data than the labeled data. Our trained networks may not be capable of detecting OOD objects populating the regions of the input spaces not covered by the UL data. However, the regions that UL data encompass in the input space can be easily extended because they obviate high costs in labeling training samples.

The rest of the paper is organized as follows. In Section 2, we introduce ID, labeled OOD (LOOD), and UL data sets used for the training and inference of networks. Section 3 describes the proposed multistage-training approach for the two tasks: photometric redshift estimation and OOD detection. Section 4 focuses on the performance analysis of the networks for both tasks using ID, LOOD, and UL test data sets. Finally, we summarize our contribution and discuss the future research direction in Section 5.

2. Data

As described in Paper I, we use the photometric data retrieved from the public data release 1 of the Pan-STARRS1 (PS1) survey as input (Kaiser et al. 2010). The PS1 survey provides photometry in five grizy bands (Chambers et al. 2016). The input data comprise 17 color-related features: four colors ($g - r$, $r - i$, $i - z$, and $z - y$) in point-spread function (PSF) measurement; their uncertainties derived using the quadrature rule; the same quantities in Kron measurement; and reddening $E(B-V)$. We use only valid photometric data, which are found under the condition of $\text{ObjectQualityFlags} == \text{GOOD}$ (Flewelling et al. 2020). To make each input feature contribute a 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 data sets in training and validating our model: the ID, UL, and LOOD data sets. Each data set is used for a different purpose. The ID data are galaxy samples used as training data for photometric redshift estimation. We use 1,480,262 galaxies, which are identical to the samples used in Paper I, as summarized in Table 1. The dominant fraction of the training samples comes from the Sloan Digital Sky Survey (SDSS) data set, where about half of them are brighter than 19 mag in r band (see the Appendix). We assign 80%, 10%, and 10% of the samples to the training, validation, and test sets (Murphy 2012) in estimating photometric redshifts for training models, finding the best model, and examining the found model, respectively. Because we randomly split data and have plenty of samples, it is reasonable to assume that the samples allocated to each set follow the same distribution as the ID samples.

### Table 1

Spectroscopic Galaxy Redshift Samples

| Data Set 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)                                         | Jones et al. (2009)        |
| PRIMUS              | 11012             | (CLASS == GALAXY) and (ZQUALITY == 4)                                             | Cool et al. (2013)         |
| 2dFGRS              | 70000             | (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)                                                | Scodberg 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 Fêvre 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 nlal) 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)     |

---

4 We interchangeably use OOD scoring and detection. Specifically, we use the term “scoring” when emphasizing the sequential property of the OOD score and use “detection” otherwise.
The UL data contain UL samples presumably containing both ID and OOD samples, and we do not know their physical classes and spectroscopic redshifts. We use the UL data for the unsupervised training of the model for OOD detection. We construct the entire UL data set to have 300,055,711 unknown objects following the same selection condition adopted for the ID training samples. From the UL data set, we evenly draw samples according to R.A. for unbiased training of the model for OOD detection. We use an R.A. interval of 10° to divide the UL data set and randomly choose 1,000,000 samples per R.A. interval. Hence, 36,000,000 UL samples are drawn from the entire UL data set. Then, we assign 80%, 10%, and 10% of the samples to the training, validation, and test sets for the purpose of training models, finding optimized models, and investigating the found models as described later, respectively.

The LOOD data comprise physically OOD objects; 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 DR1 following the same selection condition adopted for the ID training samples. The sources of these spectroscopic objects are summarized in Tables 2 and 3. We use the LOOD data for the quantitative assessment of OOD detection performance and exclude them from training the models. Hence, we use the entire LOOD samples as the test data for model validation.

### 3. Method

To equip the networks with OOD scoring/detection functionality, we train two neural networks $F_1$ and $F_2$ sharing the same structure in a multistage-training approach comprising supervised and unsupervised steps (hereinafter $step_{sup}$ and $step_{unsup}$, respectively; see Figure 1; Yu & Aizawa 2019). The two networks are trained differently in $step_{sup}$ to diverge their results on potential OOD samples and evaluate the OOD score based on their inference difference. For UL data, one network tries to have a peaked probabilistic distribution of inference on photometric redshifts, while the other network intends to generate flat probabilistic distribution as explained later in Section 3.2. The ID data keep the two networks from diverging on ID samples since the training loss using ID samples fosters the two networks to have consistent peaked distributions around the true class, i.e., correct redshift bin.

The training approach consists of three stages: supervised pre-training (training stage 1) with the labeled spectroscopic training samples, iterative supervised and unsupervised training for OOD scoring/detection (training stage 2) with the labeled spectroscopic samples and UL data, respectively, and supervised training for photometric redshift estimation (training stage 3) as stage 1 conducts. The first two stages of the training procedures are originally proposed in Yu & Aizawa (2019), whereas the third stage is added to improve the model performance for the original task, i.e., photometric redshift estimation, with the added networks for the OOD detection.

We adopt the same supervised learning technique as Paper I in estimating photometric redshifts as $step_{sup}$ and the unsupervised method using two networks introduced by Yu & Aizawa (2019) in OOD detection as $step_{unsup}$. In the following subsections, we outline $step_{sup}$, $step_{unsup}$, the three-stage training procedure, and assessment metrics to measure model performance. For a more detailed explanation of each training step, refer to Paper I and Yu & Aizawa (2019).

#### 3.1. Supervised Training Step for Photometric Redshift Estimation

In $step_{sup}$ using only ID data, we bin redshift ranges and classify samples into the redshift bins instead of conducting typical regression of redshifts as we name the model the multiple-bin regression with the neural network used in Paper I. We adopt 64 independent bins with a constant width following the performance result in Paper I. Then, the model estimates

| Data Set 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 bln or bal) and (Z_CONF == 4) | Trump et al. (2009) |
| 6dFGS         | 49                | (QUALITY_CODE == 6) or (REDSHIFT > 1.0) | 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) |

| Data Set 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) |
probabilities that the photometric redshifts of input samples lie in each bin. Using the model output probabilities, we may obtain point estimation of photometric redshift \( z_{\text{phot}} \) by averaging the central redshift values of the bins with the output probabilities.

As a training loss of step \(_{\text{sup}}\), we use anchor loss \( L_{\text{ANCH}} \). The loss is designed to measure the difference between two given probability distributions considering prediction difficulties due to various reasons, e.g., the lack of data or the similarities among samples belonging to different classes, i.e., redshift bins. The anchor loss assigns the difficult samples for prediction high weights governed by a weighting parameter \( \gamma \). For an in-depth definition of the loss, refer to Ryou et al. (2019). Since we train two networks with the same architecture for the dual tasks, the training loss of step \(_{\text{sup}}\), which should be minimized, is set as follows:

\[
L_{\text{sup}} = L_{\text{ANCH}}(p_1(z|x_{\text{ID}})) + L_{\text{ANCH}}(p_2(z|x_{\text{ID}})),
\]

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

### 3.2. Unsupervised Training Step for Out-of-distribution Detection

step\(_{\text{unsup}}\) uses the UL data, presumably containing both ID and OOD samples, in training models and employs the disparity of the results from two networks to evaluate the OOD score of the samples. The two networks are trained to maximize disagreement between them in the photometric redshift estimation for the potential OOD samples, pushing the OOD samples outside the manifold of the ID samples. The two networks, which are identically structured and trained on the same ID data via supervised learning, might have different results on OOD samples because of stochastic effects, although the networks are not optimized for OOD detection. Then, by defining discrepancy loss \( L_{\text{DCP}} \) to measure the disagreement and maximizing it, the two networks can be forced to produce divergent results for the potential OOD samples, which generally correspond to the samples in the class (i.e., redshift bin) boundaries and/or are misclassified (i.e., assigned to an incorrect redshift bin). The \( L_{\text{DCP}} \) defined for this purpose is as follows:

\[
L_{\text{DCP}}(p_1, p_2) = H(p_1(z|x)) - H(p_2(z|x)),
\]

where \( H(\cdot) \) is the entropy for given probability distributions. As the networks are trained to maximize the discrepancy \( L_{\text{DCP}} \), the entropies from \( F_1 \) and \( F_2 \) output increase and decrease, respectively. The network \( F_1 \) outputs a flat probability distribution (i.e., high entropy, high uncertainty), and the network \( F_2 \) generates a peaked one (i.e., low entropy, low uncertainty). Notably, the \( L_{\text{DCP}} \) is available in the unsupervised approach using the UL samples since it can be computed without sample labels (i.e., redshifts). The potential ID samples make it difficult for the two networks to produce discrepant inference results for photometric redshifts since the inference for the ID samples has a strongly peaked distribution in both networks \( F_1 \) and \( F_2 \) as explained below. Meanwhile, the
potentially OOD samples can increase $L_{DCP}$, as the two networks become easily divergent in the training process.

To prevent divergent results on the ID samples, we employ the sum of $L_{sup}$ and $L_{DCP}$ as the training loss of $step_{unsup}$. Although the unsupervised approach using the $L_{DCP}$ diverges the model outputs on the probable OOD examples included in the UL data set, it can also cause discrepant results on the ID samples since the UL data include ID and OOD samples. Hence, defining the training loss of $step_{unsup}$ as the sum helps the two network outputs on the ID samples stay as similar as possible since minimizing $L_{sup}$ makes the network outputs have unimodal probability distributions with a peak at the true class on ID samples. Namely, $step_{unsup}$ diverges model outputs for OOD samples more than ID ones because OOD samples are out of support of the $L_{sup}$. The training loss of $step_{unsup}$, which should be minimized in training, is defined as follows:

$$L_{unsup} = L_{sup} + \max(m - L_{DCP}(p_1(z|x_{UL}), p_2(z|x_{UL})), 0),$$

where $x_{UL}$ is an UL input vector and $m$ is a margin. Notably, $L_{DCP}$ is maximized, as $L_{unsup}$ is minimized owing to the negative sign. The use of the margin $m$ is required to prevent overfitting of the networks by replacing the second term of the unsupervised loss with zero if $L_{DCP} > m$. We use $m = 4.3$, found as the best parameter in multiple trial runs.

Our usage of $L_{unsup}$ as the combination of $L_{sup}$ and $L_{DCP}$ can be understood as multitask learning of both estimating photometric redshifts and evaluating OOD scores. Particularly, this multitask learning model is a case of transferring the knowledge of supervised learning (i.e., the model with $L_{sup}$ in estimating photometric redshifts) to unsupervised learning (i.e., the model with $L_{unsup}$ in evaluating OOD scores; Pentina & Lampert 2017). The combined loss in our dual-task learning adopts the parameter $m$, which regularizes the contribution of $L_{DCP}$ to the total loss $L_{unsup}$. This is a simple strategy to find balance among multiple tasks as commonly used as weight parameters with multiple losses in multitask learning (Sener & Koltun 2018; Gong et al. 2019; Vandenbende et al. 2021).

### 3.3. Three-stage Training

We proceed with the three-stage training of the two networks. Each training stage comprises the aforementioned $step_{sup}$ and $step_{unsup}$ as summarized in Figure 1. Training stage 1 is a pre-training of the two networks with the $step_{sup}$. The purpose of this stage is to obtain a stable performance of the networks in estimating photometric redshifts. In training stage 2, the two networks are trained for OOD detection using an iterative training procedure comprising one $step_{sup}$ and two $step_{unsup}$. Although $L_{sup}$ included in $L_{unsup}$ helps the disparity between two network outputs on ID samples stay small, the disparity increases as the models are trained by the optimization using UL samples and $L_{DCP}$. Composing stage 2 as the iterative training procedure incorporating $step_{sup}$ further precludes the disparity on the ID samples from increasing, as $L_{sup}$ included in $L_{unsup}$ affects the total loss (see Equation (3)). Therefore, as the training continues, the disagreement of the network outcomes for the probable OOD samples becomes larger than that for the ID samples because the OOD samples are outside the support of the $step_{sup}$ which is only applied to the ID galaxies. Then, the measure of disagreement (i.e., $L_{DCP}$)

![Figure 2](image.png) Comparison between spectroscopic and photometric redshifts from the high-entropy ($F_1$; left) and low-entropy ($F_2$; right) networks in training stage 2. The scatters are color-coded according to number density. The dashed lines correspond to slope-one lines and catastrophic error boundaries, respectively.
can be used as an OOD score to flag OOD candidates. At the inference level, the trained networks in training stage 2 are used to score the test samples as OOD.

In addition to training stages 1 and 2, we proceed to an additional supervised training stage, training stage 3, with step\textsubscript{sup} for photometric redshift estimation. The unsupervised training for OOD detection affects the model performance of the original task (i.e., photometric redshift estimation) because the model in training stage 2 is optimized for two losses described in Equation (3). In practice, training stage 3 might not be needed if the models trained in stage 1 are saved and used for the task of photometric redshift estimation. In this work, we train the two networks again using only \textit{L}\textsubscript{sup} after training stage 2. Then, the outputs of the two stage 3 networks are combined by averaging them to estimate photometric redshifts.

### 3.4. Assessment Metrics

Since our model performs the two tasks together, we need separate metrics to assess model performances on both tasks. When using a mean value with probabilistic inference by the trained model as the point estimation of photometric redshifts, we adopt the ones used in Paper I: bias, MAD, \( \sigma \), \( \sigma_{68} \), NMAD, and \( R_{\text{cat}} \). For a detailed explanation of the metrics, refer to Paper I. In short, the bias is measured as the absolute mean of \( \Delta z \), which is the difference between photometric redshifts and true spectroscopic redshifts. Similarly, the MAD is the mean of the absolute difference \( |\Delta z| \), \( \sigma \) and \( \sigma_{68} \) correspond to the standard deviation of \( \Delta z \) and the 68th percentile of \( |\Delta z| \), respectively. NMAD is defined to be \( 1.4826 \times \text{the median of } |\Delta z| \). \( R_{\text{cat}} \) presents the fraction of data with \( |\Delta z| > 0.15 \). Notably, the lower the point-estimation metrics are, the higher the performance is.

Before explaining OOD detection metrics, we first define four measures: true negative (TN), true positive (TP), false negative (FN), and false positive (FP). The Boolean values (i.e., true and false) in the terms indicate whether the predicted class of the sample is correct or incorrect. The negative and positive in the terms mean the true class in which the given sample is included. We set ID and OOD as negative and positive, respectively, since the model performs OOD detection as a task. Hence, for example, the number of TP cases is the number of correctly classified OOD samples. The numbers of these four cases can be varied with respect to the adopted threshold of the OOD score. Notably, these metrics are used with the LOOD samples not in training the model but in validating the trained model.

We use the central value between the minimum and maximum \( L_{\text{DCP}} \) as the threshold to compute the metrics. Note that this threshold is chosen simply because it can be easily estimated for model assessment. The following is a brief explanation of the detection metrics:

1. **Accuracy**: ratio of the correctly classified samples to the entire samples \( \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \).
2. **Precision**: ratio of the correctly classified positive samples to the entire positively classified samples \( \frac{\text{TP}}{\text{TP} + \text{FP}} \).
3. **True Positive Rate (TPR) or recall**: fraction between the correctly predicted positive samples and the entire positive samples \( \frac{\text{TP}}{\text{TP} + \text{FN}} \).
4. **\( F_\beta \)** harmonic mean of precision and recall weighted by \( \beta \) \( \frac{(1 + \beta^2) \times \text{Precision} \times \text{Recall}}{\beta \times \text{Precision} + \text{Recall}} \). The lesser and larger \( \beta \) give more weight to precision and recall, respectively. Typical values for \( \beta \) are 0.5, 1.0, and 1.5. In this work, we set \( \beta = 1.5 \) (\( F_{1.5} \)) to focus on more recall than precision, paying attention to the number of the correctly classified OOD samples among all OOD samples.
5. **AUCROC**: area under the receiver operating characteristic (ROC) curve. The ROC curve is a visualized measure of the detection performance for the binary classifier. In the curve, the $TPR$ is plotted as a function of the FP rate for different thresholds of the OOD score. FP rate is a fraction between the incorrectly predicted negative samples and entire negative samples $\frac{FP}{TN + FP}$. The $AUC_{ROC}$ offers a comprehensive measure of the detection performance of the model.

6. **AUCPR**: area under the precision-recall (PR) curve. Although $AUC_{ROC}$ offers a representative measure of binary detection performance generally, it may produce misleading results for imbalanced data because $TPR$ (i.e., the vertical axis of the ROC curve) depends only on positive cases without considering negative cases. In the PR curve, precision is plotted as a function of recall. Because precision considers both positive and negative samples, $AUC_{PR}$ may be more informative for imbalanced data.

The higher values of the detection metrics indicate better detection performance of the model. Notably, these OOD detection metrics are evaluated with the labeled ID (i.e., galaxies) and OOD (i.e., stars and QSOs confirmed spectroscopically).

### 4. Results

We present the results of the model in testing the two tasks. All results in this section are produced with the samples in the test data set unless otherwise stated.
4.1. Photometric Redshift Estimation on In-distribution Samples

We find that the networks in training stage 3 outperform the networks in stage 2 in terms of photometric redshift estimation. To verify the effectiveness of the additional training stage 3, we compare the quality of the photometric redshifts derived from the models in training stages 2 and 3. Table 4 shows the point-estimation metrics evaluated in the high- and low-entropy models (i.e., $F_1$ and $F_2$, respectively) in training stage 2 and the averaged model in stage 3. Except for the bias, the training stage 3 model outperforms others. It proves that unsupervised training of the networks in training stage 2 for OOD detection affects the quality of photometric redshifts as expected for multitask learning.

The networks in training stage 2 hold noticeable peculiarities showing deteriorated photometric redshifts. As shown in Figure 2, the photometric redshifts from the high-entropy model are overestimated for a large fraction of samples. Meanwhile, the photometric redshifts from the low-entropy model show concentrations at $z_{\text{phot}} \sim 0.05$ and $z_{\text{phot}} \sim 0.2$. The peculiarities in the derived photometric redshifts arise from the minimization of $L_{\text{unsup}}$ during training stage 2, which makes the output probability distributions of the high- and low-entropy models flat and peaked, respectively. These oddities of the training stage 2 model vanish in the averaged stage 3 model (see Figure 3). The Pearson correlation coefficient $\rho = 0.959$ affirms that the estimated photometric redshifts match spectroscopic redshifts well.

4.2. Out-of-distribution Score for Labeled Data

As we design and expect, the training stage 2 model shows better separation between the ID and OOD samples (i.e., physically OOD samples of stars and QSOs) than the stage 3 model. We compare the OOD detection performance of the networks in training stages 2 and 3. The detection metrics of the networks in training stages 2 and 3 are tabulated in Table 5, showing that the OOD detection model is optimized in training stage 2.

Interestingly, the training stage 3 model has higher precision than the stage 2 model, which is due to the high TP values with the small number of samples flagged as OOD by the training stage 3 model; 352,055 samples ($\sim 7\%$) among 5,006,223 true OODs are correctly classified as OOD. This small number results in unexpectedly high precision, AUC$_{\text{ROC}}$, and AUC$_{\text{PR}}$ for the training stage 3 model.

The stage 2 model separates well the labeled ID and OOD samples as shown in Figure 4, displaying the $L_{\text{DCP}}$ distributions of the ID and OOD samples. The $L_{\text{DCP}}$ ranges, the same ranges as used in Figures 5 and 7. Meanwhile, the photometric redshifts from the low-entropy model show concentrations at $z_{\text{phot}} \sim 0.05$ and $z_{\text{phot}} \sim 0.2$. The peculiarities in the derived photometric redshifts arise from the minimization of $L_{\text{unsup}}$ during training stage 2, which makes the output probability distributions of the high- and low-entropy models flat and peaked, respectively. These oddities of the training stage 2 model vanish in the averaged stage 3 model (see Figure 3). The Pearson correlation coefficient $\rho = 0.959$ affirms that the estimated photometric redshifts match spectroscopic redshifts well.

4.2. Out-of-distribution Score for Labeled Data

As we design and expect, the training stage 2 model shows better separation between the ID and OOD samples (i.e., physically OOD samples of stars and QSOs) than the stage 3 model. We compare the OOD detection performance of the networks in training stages 2 and 3. The detection metrics of the networks in training stages 2 and 3 are tabulated in Table 5, showing that the OOD detection model is optimized in training stage 2.

Interestingly, the training stage 3 model has higher precision than the stage 2 model, which is due to the high TP values with the small number of samples flagged as OOD by the training stage 3 model; 352,055 samples ($\sim 7\%$) among 5,006,223 true OODs are correctly classified as OOD. This small number results in unexpectedly high precision, AUC$_{\text{ROC}}$, and AUC$_{\text{PR}}$ for the training stage 3 model.

The stage 2 model separates well the labeled ID and OOD samples as shown in Figure 4, displaying the $L_{\text{DCP}}$ distributions of the ID and OOD samples. The $L_{\text{DCP}}$ ranges, the same ranges as used in Figures 5 and 7.
samples. The distribution of the $L_{DCP}$ in the UL samples makes sense because the largest fraction of sources should be stars in our Galaxy.

The prediction difficulties of photometric redshift estimation and OOD scores for the ID data are related as presented in Figure 5. Although most samples with noncatastrophic (hereinafter noncat) estimation (i.e., $(z_{\text{spec}} - z_{\text{phot}})/(1 + z_{\text{spec}}) \leq 0.15$) are assigned to the low-$L_{DCP}$ group, most samples with catastrophic (hereinafter cat) estimation of photometric redshifts belong to middle- or high-$L_{DCP}$ groups. High-$L_{DCP}$ samples generally deviate more from the slope-one line than low-$L_{DCP}$ samples, indicating that the samples with high OOD score can be viewed as OOD-like ID samples and tend to have incorrect photometric redshifts.

Figure 9. $L_{DCP}$ distributions of QSOs (top) and stars (bottom) with respect to the input colors in PSF measurement. The black contour lines depict the ID training sample distribution. The points are discretely color-coded according to $L_{DCP}$ groups. The samples of stars used to plot here are randomly undersampled approximately by a factor of 10 from the entire star LOOD data set.
The distributions of LDCP for the ID noncat and cat samples shown in Figure 6 also affirm our interpretation of the relation between LDCP and the error in photometric redshifts. The distributions show that a higher fraction of cat samples populate in the high LDCP ranges than noncat samples. Although the number of noncat samples decreases as the LDCP increases, the cat samples are almost uniformly distributed over the entire LDCP range. The difficulties in estimating photometric redshifts are related to the OOD-like ID samples.

The training stage 3 model is unreliable concerning cat samples with high LDCP. As shown in Figure 6, the number of noncat samples with high confidence on photometric redshifts decreases as LDCP increases. The cat samples mostly have low confidence on photometric redshifts throughout low and middle LDCP ranges. It also shows that the networks in training stage 3 are well calibrated in estimating photometric redshifts for the cat samples, with low and middle LDCP presenting the low confidence on them. However, the cat samples with high LDCP extend to cases with high confidence, which is an indication of overconfident results on the wrong photometric redshift estimation for OOD-like ID cat samples. It also emphasizes the importance of the OOD detection algorithm since we cannot filter the OOD samples simply based on confidence on photometric redshifts without the OOD score.

The distribution of LDCP in the space of input features allows us to visually inspect the OOD-like ID samples. In Figure 7, most high-LDCP samples reside in the low-density regions of the training samples. The training stage 2 model is optimized to assign high OOD scores to underrepresented galaxies, meaning that these high-LDCP ID samples are underrepresented galaxies from the model perspective.

As we find the ID samples with high LDCP, there are LOOD samples with low LDCP (i.e., the ID-like LOOD samples). As presented in Figure 8, depicting the LDCP distributions for the two types of LOOD objects, i.e., QSOs and stars, a fraction of LOOD samples have low LDCP, although most samples have high LDCP. The points to be noted are as follows: (1) both QSO and star distributions have peaks at the lowest LDCP bin, (2) the ratio of the lowest LDCP peak to the highest LDCP peak for QSOs is larger than that of stars, and (3) more QSOs are distributed in the low and middle LDCP ranges than stars.

The ID-like LOOD samples locate around the ID sample distribution in the input space. Figure 9 shows that the QSO samples with high LDCP overwhelm the ones with low LDCP outside the ID sample density contours. However, the number of the low-LDCP samples for QSOs increases as they approach the ID sample density peak. These ID-like QSOs contribute to the peak at the lowest LDCP bin of Figure 8.

As found in Figure 9, some stars certainly reside inside the contours of the ID galaxy samples with respect to the projected input dimension, corresponding to the stars with low LDCP (Figure 8). However, most stars have input features significantly deviating from the typical properties of the ID galaxy.

---

5 In Paper I, we have already found that the lack of training samples causes high errors in photometric redshifts.

6 This visual presentation often used hereafter displays the projected dimensions of the input space with degeneracy of the input features in the higher dimension.
samples in any one dimension of the input features. Compared with the QSO distribution shown in Figure 9, more stars than QSOs can be sorted out from the ID samples in higher dimensions of input features since the star distribution significantly differs from that of the ID samples. Therefore, the relatively smaller peak of stars at the lowest OOD score and a lower fraction of ID-like stars than QSOs arise as shown in Figure 8.

Filtering out the QSO and star samples with low/middle $L_{DCP}$ enables further reliable usage of the trained model for photometric redshifts. As depicted in Figure 10, although the confidence of ID samples gradually reduces as $L_{DCP}$ increases, the LOOD objects with high $L_{DCP}$ have higher confidence than those with low $L_{DCP}$. Using this distribution difference in the space of $L_{DCP}$ and confidence, a considerable number of LOOD samples with low and middle $L_{DCP}$ can be filtered by checking their higher confidence than that of ID samples.

4.3. Out-of-distribution Score and Photometric Redshifts of Unlabeled Data

Because the UL data have no labels and spectroscopic redshifts, we assess the model performance on the UL samples indirectly. To measure the performance of detecting the physically OOD samples in the UL data set, we first extract the point-source score (hereinafter ps-score) from Tachibana & Miller (2018) for the UL samples. The ps-score is a machine-learning-based probabilistic inference on being point sources for PS1 objects with the range of $[0,1]$. The samples with low ps-score are more likely to be extended sources, i.e., mainly galaxies, and we expect our OOD detection algorithm to produce low $L_{DCP}$ for samples with low ps-score.

The distributions of $L_{DCP}$ with respect to ps-score are subject to the expected pattern. Figure 11 shows the $L_{DCP}$ distribution of the samples undersampled from the UL data set within the R.A. ranges $[0,10]$, $[100,110]$, and $[170,180]$. 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 shift toward high $L_{DCP}$ ranges. We conjecture that the middle/high-$L_{DCP}$ samples with low ps-score are highly likely to be underrepresented galaxies residing in the outskirts or low-density parts of the ID distribution (see discussion in the Appendix). Except for this understandable difference, the two machine-learning models reach a consensus on other UL samples that are mainly point-source stars and QSOs.

Figure 12 indirectly proves $L_{DCP}$ to be a reliable OOD score for a broad range of input space for the UL samples. The UL samples with low $L_{DCP}$ have the most similar distributions to those of the ID samples because the networks assign low $L_{DCP}$ to the UL samples with input features comparable with the ID samples. The UL samples with middle $L_{DCP}$ begin to display discrepancies from the ID samples in input feature distributions. The samples in the high-$L_{DCP}$ group have a completely different distribution from the ID samples. In the increasing order of the OOD score (i.e., $L_{DCP}$), the UL data density distributions in the input color spaces deviate more from those of ID samples. As expected from the typical colors of stars (e.g., Jin et al. 2019), the peak in the color distribution for the high-$L_{DCP}$ UL samples in Figure 12 corresponds to common stellar colors.

The training stage 3 model adequately produces photometric redshifts for the UL samples. As examined with the ID and LOOD samples, we expect the photometric redshift distributions to be similar to those of ID samples for the UL samples with low $L_{DCP}$ and random guessing, i.e., uniform distribution, for the UL samples with high $L_{DCP}$, respectively. As presented in Figure 13, the samples in the low $L_{DCP}$ range mostly reside inside the ID sample density contours. On the other hand, the photometric redshifts of the OOD-like samples randomly spread in the spaces without any strong systematic patterns. It certainly shows that the photometric redshift distributions for OOD-like samples are closer to a uniform distribution than ID-like samples. However, most redshifts are in the region $z_{phot} < 0.8$. This arises from the fact that approximately 99.82% of the training samples for photometric redshifts are found in the region $z_{spec} < 0.8$. 

Figure 12. Density contour lines of the UL samples with filled density contour map of the ID samples in the space of input colors ($g − r$) and ($r − i$) (PSF measurement). We use the three $L_{DCP}$ ranges to split the UL samples into low-, middle-, and high-$L_{DCP}$ groups as shown in Figure 5.
We examine the UL samples with the low ps-score and their \( L_{\text{DCP}} \) in the input feature space. These samples generally correspond to galaxies not well represented by the training data. Naturally, the UL samples include more faint galaxies than the spectroscopic training samples because the procedure of selecting spectroscopic targets is generally biased to bright samples. Furthermore, the color and photometric redshift distributions of the UL samples with the low ps-score and high \( L_{\text{DCP}} \) affirm that these UL samples have input features that are not well represented by the training samples (see Figure 14).

5. Discussion and Conclusion

In this study, we propose a new approach not only to estimate photometric redshifts but also to check the reliability of such redshifts through OOD scoring/detection. The proposed model successfully estimates the photometric redshifts of galaxies using our previous model (presented in Paper I), even with an additional network to detect the OOD samples. Our trained model is available with the Python code, which is based on PyTorch (Paszke et al. 2019), on GitHub\(^7\) under an MIT License (Lee & Shin 2021) for public usage.

The proposed method to detect the OOD data with the photometric redshift inference model shows that the new model can measure how much OOD data deviate from the training data. Most samples with a catastrophic redshift estimation correspond to data with high OOD scores (see Figure 5). However, our model estimating the OOD score cannot replace models that separate galaxies from stars and QSOs. As shown in Figure 10, numerous stars and QSOs exhibit low OOD scores and high confidence values for photometric redshifts in

\(^7\)https://github.com/GooLee0123/MBRNN_OOD

---

**Figure 13.** \( L_{\text{DCP}} \) distributions of the UL samples in the space of input colors (PSF measurement) and estimated photometric redshifts. The black contour lines are drawn with the spectroscopic redshifts and corresponding input colors of the ID samples. The samples used to plot here are randomly undersampled by a factor of 10 from the entire UL data set.

**Figure 14.** \( L_{\text{DCP}} \) distributions of the low ps-score (<0.1) UL samples in the space of input colors (PSF measurement) and estimated photometric redshifts. The black contour lines are drawn with the spectroscopic redshifts and corresponding input colors of the ID samples.
our model, although they are physically OOD objects. These objects have indistinguishable input features compared with galaxies in terms of the features used for photometric redshift estimation.

We plan to apply the SED-fitting methods to galaxies with high OOD scores or low confidence values. The machine-learning inference model has limitations that depend on the training samples. Therefore, combining the SED-fitting methods with machine-learning inference models can be an effective strategy to achieve high analysis speed and accuracy. This would have broad applicability in estimating photometric redshifts for numerous galaxies expected for future surveys such as the Legacy Survey of Space and Time (Ivezić et al. 2019).

The current implementation using two networks needs to be improved to see the full benefits of ensemble learning. The multiple anchor loss parameters presented in Paper I are required for more precise estimation of photometric redshifts and evaluation of the OOD scores in a single machine-learning framework. The proposed model in this second paper of the series excludes multiple inferences of photometric redshifts in ensemble learning. The ensemble model distillation might be a possible technique 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 underrepresented galaxies in the 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, thereby reducing the fraction of incorrect estimation and/or the uncertainty of estimation (Charpiat et al. 2019; Pruthi et al. 2020). Because the influence of data relies on the model, the future method needs to have a model-dependent algorithm to assess the influence of such data quantitatively. The highly influential data among the OOD samples require 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 underrepresented galaxies will be also crucial in improving the evaluation of the OOD score and finding useful samples that need spectroscopic labeling. A large fraction of OOD samples might be produced by effects such as source blending (including physically merging galaxies) and bias in the spectroscopic target selection. When selecting highly influential OOD samples among the UL data, our new algorithm for the OOD score may need a step of filtering out the photometrically unreliable samples such as blended galaxies. The underrepresented galaxies might include galaxies with less common physical properties, e.g., extreme dust obstruction, unusual stellar populations, short post-starburst phase, unusually intensively star-forming galaxies, merging galaxies at various evolutionary stages, and galaxies with weak QSO emission.

Semisupervised learning methods might be an alternative approach to supervised learning methods such as our proposed approach. Although semisupervised learning has its own limitations, such as smoothness assumption (Chapelle et al. 2006), semisupervised learning methods intrinsically do not have the OOD problem, except for physically OOD samples. However, the efficacy and reliability of the semisupervised learning is a new challenge compared with supervised learning methods. Measuring the influence of data on the training and implementation of the algorithm is crucial because the performance of semisupervised learning methods is strongly subjective to the labeled samples and the weights of UL samples.

We thank David Parkinson for his careful reading and thoughtful comments. We are also grateful to Kyungmin Kim and Hyung Mok Lee for insightful discussions. We also thank the anonymous referee for helpful comments with valuable suggestions. This research has made use of the VizieR catalog access tool, CDS, Strasbourg, France (DOI: 10.26093/cds/vizier). The original description of the VizieR service was published in 2000, A&AS 143, 23. The Pan-STARRS1 Surveys (PS1) and the PS1 public science archive have been made possible through contributions by the Institute for Astronomy, the University of Hawaii, the Pan-STARRS Project Office, the Max-Planck Society and its participating institutes, the Max Planck Institute for Astronomy, Heidelberg and the Max Planck Institute for Extraterrestrial Physics, Garching, Johns Hopkins University, Durham University, the University of Edinburgh, the Queen’s University Belfast, the Harvard-Smithsonian Center for Astrophysics, the Las Cumbres Observatory Global Telescope Network Incorporated, the National Central University of Taiwan, the Space Telescope Science Institute, the National Aeronautics and Space Administration under grant No. NNX08AR22G issued through the Planetary Science Division of the NASA Science Mission Directorate, the National Science Foundation grant No. AST-1238877, the University of Maryland, Eotvos Lorand University (ELTE), the Los Alamos National Laboratory, and the Gordon and Betty Moore Foundation.

Appendix

Example of Underrepresented Galaxies in Training Data and \( L_{\text{DCP}} \)

We demonstrate a typical case of selection effects based on bright objects in spectroscopic training data as an example of considering OOD data. Since the brightness of objects is a main factor affecting target selection in spectroscopy, we examine an extreme case when training samples of spectroscopic redshifts consist of only bright objects, and data for inference include only faint objects. This case undoubtedly demonstrates that our model to detect OOD data works well.

Adopting the neural network model in the same way as described in Section 3, we train the model with the bright SDSS spectroscopic objects, which have \( r \)-band Kron magnitude less than or equal to 19, as supervised training samples. The properties of the 634,928 bright SDSS objects are different from those of the 659,114 faint objects as expected in the SDSS data (see Figure A1). The bright objects generally have lower redshifts and bluer color than the faint objects have. Therefore, a large fraction of the faint objects should be OOD objects with respect to the trained model as underrepresented galaxies in the training data.

The performance in estimating photometric redshifts is much worse for the faint objects than for the bright test samples. As shown in Figure A2, the trained model cannot estimate correct photometric redshifts for the majority of the faint objects because they are underrepresented objects, i.e., OOD objects. The point-estimation metrics also show that the model is not
able to infer photometric redshifts for the faint objects. For example, $R_{\text{cat}}$ values are 0.0012 and 0.22 for the bright test samples and the faint objects, respectively. The bias value for the bright test samples is about 0.004, while that value for the faint objects is about 0.09.

Our method to evaluate $L_{\text{DCP}}$ works well in this demonstration. Figure A3 shows that our method successfully gives the faint objects high $L_{\text{DCP}}$. The training data composed of the bright objects do not cover a substantial range of redshifts and input features for the faint test objects as presented in Figure A1. Most of the faint objects have high $L_{\text{DCP}}$ as we expect. Because most UL samples are stars in our Galaxy or faint objects, they also show high $L_{\text{DCP}}$ values in Figure A3.

Colors, as well as their high uncertainties of the faint SDSS objects, which have the high $L_{\text{DCP}}$ values, display distinct deviation from those of the bright spectroscopic samples (see Figure A4). Because we use $r$-band magnitude to divide the SDSS spectroscopic samples into the bright and faint sets, redder objects are more likely to be OOD objects than faint objects. As presented in Figure A4, the faint spectroscopic galaxies showing high $L_{\text{DCP}}$ have significantly redder $(i - z)$ and $(z - y)$ colors than the training samples used to derive the model.

The colors of the UL samples with high $L_{\text{DCP}}$ are clearly different from those of the faint galaxies with high $L_{\text{DCP}}$. In Figure A4, the peak of the color distribution for the UL samples
with high $L_{DCP}$ is much bluer than those of both the bright and faint SDSS galaxies with high $L_{DCP}$. Considering the typical distribution of colors for stars, QSOs, and galaxies in the SDSS filter bands (e.g., Finlator et al. 2000; Hansson et al. 2012; Jin et al. 2019), the peak in the color distribution for the UL samples corresponds to the expected colors of stars and low-redshift QSOs.

**ORCID iDs**

Joongoo Lee @ https://orcid.org/0000-0003-2510-4132
Min-Su Shin @ https://orcid.org/0000-0002-9934-3139

**References**

Aguado, D. S., Ahumada, R., Almeida, A., et al. 2019, ApJS, 240, 23
Amodei, D., Olah, C., Steinhardt, J., et al. 2016, arXiv:1606.06565
Beck, R., Dobos, L., Budavári, T., Szalay, A. S., & Csabai, I. 2016, MNRAS, 460, 1371
Blake, C., & Bridle, S. 2005, MNRAS, 363, 1329
Bolzonella, M., Miralles, J. M., & Pelló, R. 2000, A&A, 363, 476
Bordoloi, R., Lilly, S. J., & Amara, A. 2010, MNRAS, 406, 881
Carliles, S., Budavári, T., Heinis, S., Priebe, C., & Szalay, A. S. 2010, ApJ, 712, 511
Canuana, R., & Niculescu-Mizil, A. 2006, in Proc. 23rd Int. Conf. on Machine Learning, ICML ’06 (New York: Association for Computing Machinery), 161
Cavuoti, S., Amaro, V., Brescia, M., et al. 2017, MNRAS, 465, 1959
Chambers, K. C., Magnier, E. A., Metcalfe, N., et al. 2016, arXiv:1612.05560
Chapelle, O., Schölkopf, B., & Zien, A. 2006, Semi-supervised Learning (Adaptive Computation and Machine Learning) (Cambridge, MA: The MIT Press)
Chaplot, G., Girard, N., Felardos, L., & Tanabalka, Y. 2019, in Advances in Neural Information Processing Systems 32: Annual Conf. on Neural Information Processing Systems, ed. H. M. Wallach et al. (Trier: DBLP), 5343, https://proceedings.neurips.cc/paper/2019/hash/c6f1571dbd2f949d3fe5ae1608dd48b-Abstract.html
Childress, M. J., Lidman, C., Davis, T. M., et al. 2017, MNRAS, 472, 273
