The Astrophysical Journal, 983:173 (10pp), 2025 April 20 https://doi.org/10.3847/1538-4357/adbe62

Identifying Catastrophic Outlier Photometric Redshift Estimates in the COSMOS Field with Machine Learning Methods

Mitchell T. Dennis, Esther M. Hu, and Lennox L. Cowie
Institute for Astronomy, University of Hawai'i at Mānoa, 2680 Woodlawn Drive, Honolulu, HI 96822, USA
Received 2023 April 28; revised 2025 March 6; accepted 2025 March 8; published 2025 April 17

Abstract

We present the result of two binary classifier ensembled neural networks to identify catastrophic outliers for photo-z estimates within the COSMOS field utilizing only eight and five photometric bandpasses, respectively. Our neural networks can correctly classify 55.6% and 33.3% of the true positives with few to no false positives. These methods can be used to reduce the errors caused by the errors in redshift estimates, particularly at high redshift. When applied to a larger data set with only photometric data available, our eight bandpass network increased the number of objects with a photo-z greater than five from 0.1% to 1.6%, and our five bandpass network increased the number of objects with a photo-z greater than five from 0.2% to 1.8%.

Unified Astronomy Thesaurus concepts: High-redshift galaxies (734); Galaxies (573); Neural networks (1933); Classification (1907)

1. Introduction

Current and upcoming large-scale sky surveys (e.g., HEROES, A. J. Taylor et al. 2023, COSMOS, P. Capak et al. 2007; O. Ilbert et al. 2008; C. Laigle et al. 2016; J. R. Weaver et al. 2022; GOODS, M. Dickinson et al. 2003; LSST, Ž. Ivezić et al. 2019; SDSS, M. R. Blanton et al. 2017) have or will take observations for millions of galaxies within a limited number of photometric bands. These measurements will provide researchers with redshifts for weak lensing, enabling them to study the large-scale structure of the universe, draw conclusions about the observed galaxies, and determine which ones warrant follow-up investigations using spectroscopy. Redshift is among the most important criteria, and therefore it is crucial that calculations of the redshift based on photometric data (e.g., using algorithms such as LePhare, S. Arnouts et al. 1999; O. Ilbert et al. 2006; EAZY, G. B. Brammer et al. 2008; SPIDERz, E. Jones & J. Singal 2017; and L. Dobos et al. 2012 from SDSS) be accurate.

While these algorithms have been largely successful, they are not perfect. The choice of method can influence the accuracy of the result. The standard approach is to fit observations to a representative set of templates (template fitting) and minimize χ^2 (e.g., M. Bolzonella et al. 2000). This method does not require spectroscopic information, and therefore is not limited by the need for spectroscopic observations. Other methods utilize training algorithms with known spectroscopic redshifts to map photometric data to redshifts. Both of these techniques are limited by the data that is used to create them (i.e., the template set or the training data) and can be unreliable. Other techniques such as using Bayesian statistics (e.g., N. Benítez 2000) or Monte Carlo methods (e.g., G. Rudnick et al. 2001) offer their own advantages and disadvantages. All photometric redshifts rely on the features shown on the observed spectral energy distribution (SED). For objects whose SEDs have less distinct features or whose distinct features (e.g., the Lyα break) have been substantially shifted into wavelengths that were not covered or easily observed, acquiring accurate photometric redshifts is even more challenging.

Although photometric redshift techniques have improved over time, photometric redshift catalogs continue to contain small populations of galaxies with redshifts that are substantially under- or overpredicted. Many different works (e.g., E. Jones & J. Singal 2020; J. Singal et al. 2022) have established different definitions for outliers and catastrophic outliers (COs). In this work, we investigate a specific subset of COs, deliberately focusing on COs that populate the high redshift universe. For our study, this subset includes all objects with a photometric redshift (photo-z) $0 \leqslant \text{photo-}z \leqslant 1.5$ and the spectroscopic redshift (spec-z) satisfies $1.8 \leqslant \text{spec-}z \leqslant 7.0$. The data in this range also satisfies the condition where the difference between the spectroscopic redshift (spec-z) and the photo-z is greater than 1.5. All other objects are designated as noncatastrophic outliers (NCOs). The range we have chosen is an important component of the analysis of the COSMOS data set (e.g., F. F. Foo et al. 2014). Objects with less distinct features in their SEDs and those with distinct features that have been substantially redshifted cause misclassification of objects by the photometric redshift algorithm. Additionally, high redshift work that currently focuses on Lyα emitters and rarer objects of other sorts, may be particularly affected by the use of SED fitting. Selectively targeting known difficult regions of parameter space with data-driven corrections will improve the quality of photometric redshifts.

As with all incorrect predictions, COs negatively impact the scientific products of these catalogs, and therefore developing algorithms and methods for identifying these outliers is crucial to achieving the scientific goals for these surveys. Rather than creating an entirely new redshift algorithm, this work focuses on positively flagging these objects for potential correction and different weighting, or even removal from large galaxy studies. Previous works in this area include E. Jones & J. Singal (2017), J. Pasquet et al. (2019), and A. Momtaz et al. (2022) using various machine learning (ML) methods, including both neural networks and support vector machines (see M. Brescia et al. 2021 for a review).

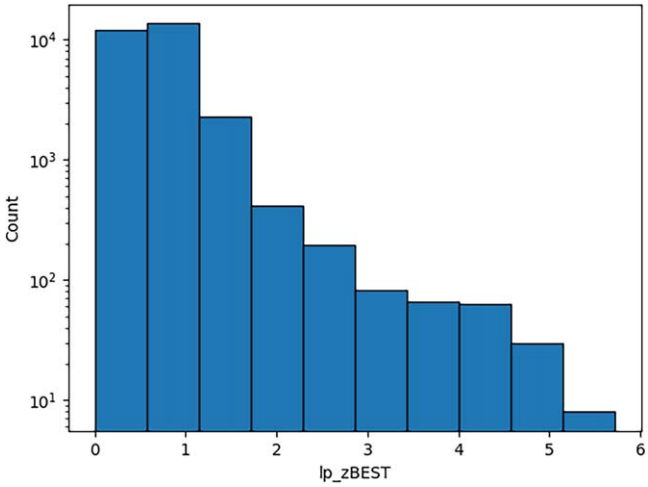

Figure 1. A histogram of roughly 29,000 photometric redshifts for the training data that is fed into the ML algorithm, presented on a logarithmic scale on the *y*-axis. The plot reveals that the majority of objects have a redshift less than 2. Within the training data, where both photo-z and spec-z are available, only a small percentage of objects have a spec-z greater than 2 (2.9%), greater than 3 (0.94%), greater than 4 (0.45%), and greater than 5 (0.10%).

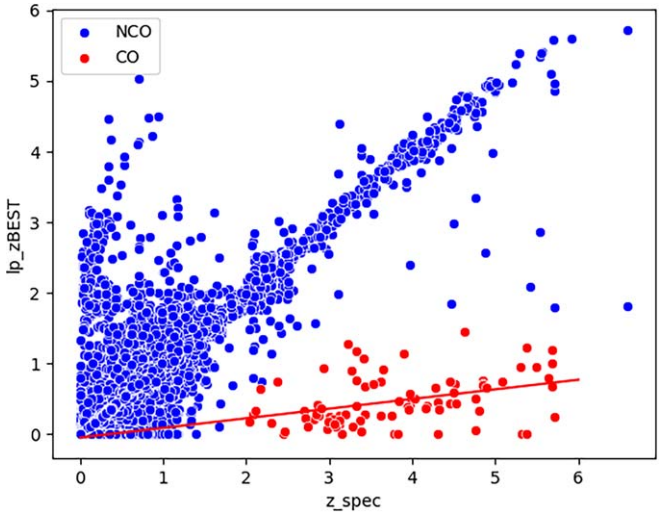

Figure 2. A plot of the training data set comparing photometric redshifts to spectroscopic redshifts. The objects highlighted in red constitute a small but crucial subset for understanding the high redshift universe. This subset includes all objects with a photometric redshift (photo-z) less than 2.0 and a difference between photo-z and spectroscopic redshift (spec-z) exceeding 1.5. Visual inspection suggests that this is a reasonable starting point for our cut. These objects represent a significant percentage of cases where spec-$z > 2$. Section 2 provides a more detailed discussion on the imbalanced nature of the data set as a whole, which may be harder to discern from this figure alone. The trend line, along with a 68% confidence interval, is utilized in Section 3 to correct flagged outliers. In Section 2, we explore how this simple correction affects the overall redshift distribution in the entire data set. The trend line has an R^2 value of 0.14, indicating that a linear relationship explains 14% of the variation between photo-z and spec-z. The p-value for the R value is small 0.0003, with a standard error of 0.036. This p-value suggests that it is highly unlikely the correlation is due to chance.

The data selection criteria are given in Section 2, a discussion of the results of ML testing is given in Section 3, an application of the network is discussed in Section 5, we investigate the source of COs in Section 4, we discuss our work in Section 6, and our conclusions are drawn in Section 7.

2. Selecting the Data

This work utilized the photo-zs provided in the COSMOS2020 photo-z catalog (J. R. Weaver et al. 2022), hereafter COSMOS2020. These data include R.A., decl., a set of photo-z solutions and their corresponding chi-squared solutions, as well as various magnitudes and their errors. The photo-zs used were calculated using the Le Phare template fitting method introduced in S. Arnouts et al. (1999) with subsequent improvements in O. Ilbert et al. (2006). Of the photometric data, only the u, B, V, g, r, i, z, and K magnitudes were used. We utilize only these few bandpasses because they provide a relatively large number of data points that have non-null values. Null values, even those documented with a representative number (e.g., 99) are problematic for ML because they introduce spurious relationships and so must be discarded. To address this, we limit our magnitude criterion to $-80 <$ABS Mag < 80, where "ABS Mag" is any of our magnitudes. Any object either completely missing photometric data or whose data indicates an observation but does not have a detection in one of these bands was dropped from the sample. Moreover, in some instances, a specific bandpass was missing but could be substituted with a similar one (e.g., the Hyper Suprime-Cam I Band magnitude was missing, but within the same data set there was a Suprime-Cam I Band magnitude). In these cases, the substitution was made in an attempt to maximize the size of our data set. Due to the similarity of the bandpasses, we do not expect this to substantially affect the ML.

Spectroscopic data were taken from the DEIMOS 10k spectroscopic sample (G. Hasinger et al. 2018), zCOSMOS spectroscopic data L. Pozzetti et al. (2010), and the current publicly available spectroscopic data from the COSMOS field (M. Salvato 2021, private communications). For these data, we restricted our study to the redshift measurements with a quality flag of greater than 1. Paraphrasing S. J. Lilly et al. (2007), of the classes we kept in our sample, class 4 are considered "completely secure redshift" with "unambiguous multiple spectral features;" class 3 are also "very secure" but with a possibility of error; class 2 are still secure, but there is a significant risk of error. These are in contrast to class 9, objects for which there is a single unambiguous emission feature, but it cannot be completely determined which feature it is between Hα, O[II], or Lyα, and so the most likely case was chosen; class 1, which is "an informed guess;" and class 0 which indicates no redshift could be calculated. Although class 2 has a significant risk of error due, we still include them because there are studies conducted which either fail to explicitly list the quality flag selection criteria (e.g., M. Jafariyazani et al. 2024; N. B. Sillassen et al. 2024; S. Thorp et al. 2024) or others that include quality flag 2 (e.g., O. R. Cooper et al. 2024), typically with an acknowledgment that the results for these objects are tentative. Our analysis focuses exclusively on objects with quality flags between 2 and 4, excluding quality flags greater than 10, which correspond to active galactic nuclei (AGN) with the 11, 12, 13, and 14 corresponding to flags 1, 2, 3, and 4, respectively. The quality flag and photometric data restrictions yielded 86 COs. Of these remaining 86, 19 have a quality flag of 4, 28 have a quality flag of 3, and 38 have a quality flag of 2 (22.1%, 32.5%, and 45.4%, respectively). A histogram of the photo-z distribution for our training data is shown in Figure 1, and a scatterplot identifying the specific subset of COs that this work focused is shown in Figure 2. Furthermore, Figure 2 shows that COs account for 14.5% of the spec-zs > 2, 24.4% of

the spec-zs > 3, 24.6% of the spec-zs > 4, and 37.9% of the spec-zs > 5. The fraction of sources with reported spectroscopic redshift after we made our photometric cuts with spec-z greater than 2, 3, 4, and 5 is 2.1%, 0.94%, 0.45%, and 0.10%. A visual inspection of Figure 2 shows a relatively clean grouping where we begin our cut, which all appear to follow a similar trend. This is in contrast to other works which typically define COs as any object where the absolute distance between the photo-z and the spec-z is greater than 1 (G. Bernstein & D. Huterer 2010; M. L. Graham et al. 2018; M. Wyatt & J. Singal 2021; J. Singal et al. 2022). Because we used the same data that produces the photo-z (and thus also produced this trend), we reasoned there was relationship in parameter space that could be exploited by ML to identify these values. The trend line is shown on the chart and has an R^2 value of 0.14, with the p-value of 0.0003 being nearly zero, i.e., it is unlikely that this trend is due to chance. There is not enough statistical confidence in the fit of this line to recover accurate photo-zs, but this toy model can give some idea of the number of very high redshift galaxies that may be currently under-reported.

Our ML network employs ensemble learning, which involves the use of numerous ML networks in parallel. Their results are combined to form a single output to maximize accuracy. Therefore, training, testing, and validation data sets were created for the ML ensemble as a whole. These splits were broken down as 80% training, 10% testing, and 10% validation. These are subsequently referred to as the ensemble data sets. We built 30 individual networks into our ensemble, and used the ensembled nature of our algorithm to account for the errors in the photometric data.

ML networks tend to perform better on data that has been normalized, and so we normalized the magnitude and error data between zero and one. Furthermore, each entry in our catalog includes error bars for each of the magnitudes and a confidence interval for the photo-zs. For an individual observation, we treated the errors in the magnitudes as standard deviations and randomly sampled Gaussian distributions centered on the ABS magnitude each of the photometric bandpasses using a standard deviation equivalent to the given error (in mag units) each respective bandpass. This procedure was repeated for each individual network in the ensemble for the training data alone. Our final ensemble network had 30 individual constituents, and therefore 30 slightly altered versions of the training data were created.

Our data set contains 28,841 NCOs and 86 COs ($\approx$0.30% COs, a highly unbalanced data set), as shown in Figure 2. ML networks perform best on balanced data sets (Y. Sui et al. 2019) (e.g., data sets which have a 1:1 ratio between different classifications). Highly unbalanced data sets are a persistent problem in ML and have been addressed by developing tools to balance them. We balanced our data by applying the synthetic minority oversampling technique (SMOTE; N. V. Chawla et al. 2011) on the individual training data sets. Instead of randomly removing data (and subsequently losing potentially valuable information), SMOTE creates synthetic data points for the minority class(es). SMOTE randomly selects a member of the minority class, calculates the convex combination between this member and its neighbors also in the minority class, and randomly selects a point within the convex combination. This process is repeated until the data set is balanced. For our data, this was only done for the training data. Unbalanced data sets will continue to pose a persistent challenge, and further tools are being developed to address this problem.

To ensure a proportional split between training, testing, and validation sets, the two classes (COs and NCOs) were separated before being randomly divided into training, testing, and validation data sets for each class (yielding six data sets). The data sets were then combined with their corresponding counterparts into the training, testing, and validation data sets.

Table 1
The Architecture of the Constituent Neural Networks that Comprise the Ensembled Network

Number	Type	Neurons/ Rate	Activation Function
1	Input	11	Linear
2	Batch Norm.	N/A	N/A
3	Dense	256	ReLU
4	Dropout	0.5	N/A
5	Dense	256	ReLU
6	Dropout	0.5	N/A
7	Dense	256	ReLU
8	Dropout	0.5	N/A
9	Dense	256	ReLU
10	Dropout	0.5	N/A
11	Dense	256	ReLU
12	Dropout	0.5	N/A
13	Dense	1	Sigmoid

Note. All densely (or fully) connected layers are given rectified linear unit activation functions except for the output neuron, which is given a sigmoid activation function. All dropout layers are given a 50% dropout rate and directly follow a dense layer (except the output layer). Only one batch normalization layer is used before the first fully connected layer directly following the input layer.

3. Creating the Network

3.1. Constructing the Network

As discussed above, our neural networks are implemented in an ensemble, meaning the results of multiple (in our case 30) neural networks are aggregated together into a single binary output (CO or NO). We implemented our neural networks using TensorFlow and its companion package Keras (M. Abadi et al. 2015; F. Chollet et al. 2015). The networks were implemented using the standard Adam optimizer (D. P. Kingma & J. Ba 2014) with additional controls, including an early stopping condition and a monitor that adjusts the learning rate. While each model is slightly different due to the random nature of each of the individual training data sets, the networks generally converge within 75 iterations (or epochs). The loss function used was binary crossentropy loss.

The hyperparameters used for the Adam optimizer were a learning rate of 0.001, $\beta_1 = 0.9$, and a $\beta_2 = 0.9$. The early stopping condition monitored the validation data loss metric (binary crossentropy) with a patience of 15 epochs, and the learning rate monitor also monitored the validation data loss metric and had a patience of 10 epochs with a factor of 0.3 reduction.

A table summarizing the architecture of the constituent networks within the larger ensemble is provided below in Table 1. The constituent networks each have 13 layers. There are 11 input columns (eight mag and three photo-z values) for

our first analysis, and eight for our second analysis (five mag and three photo-z). These inputs are immediately passed to a batch normalization layer. During training, the batch normalization layer normalizes each batch's current inputs to have a center of 0 and a standard deviation of 1. During inference mode (used for both validation and testing data), it uses a moving average and standard deviation of the batches seen during training. Following batch normalization are five dense-dropout layer combos, which all have a rectified linear unit activation function across 256 neurons and a dropout rate of 50%. The final layer is a single neuron with a sigmoid activation function that outputs a value between 0 and 1.

The results of the constituent neural networks were then ensembled together to create a final prediction. Each individual network made a prediction on the ensemble validation data. These predictions were evaluated using each individual network's custom threshold value and then rounded to 1 (CO) or 0 (NO). If a data point was evaluated by every individual network to be a CO, it would receive a combined score equal to the total number of networks.

3.2. Training and Validating the Network: Eight Bands

The input columns for our training data were the absolute magnitude columns for the u, B, V, g, r, i, z, and K bandpasses (with slight randomness based on their errors described above), and LePhare photometric redshift columns including the upper and lower 68% bounds. The spectroscopic redshift columns were not included in the ML training and were only used to classify our data as COs and NCOs.

Both the testing and validation data sets at each level are withheld from training the network. However, the validation data set was used to "guide" the network toward a solution (used as an evaluation at every iteration of the training), whereas the testing data set is a final test. After we completed the training, we evaluated the model using the validation data set, aiming to adjust the "threshold" for a positive indicator (in this case a CO). Generally, the threshold is set to 0.5 (or 50%) as in J. Singal et al. (2022). This means if an individual network outputs a value $\geqslant$0.5 for given a piece of input data, the result would be considered a positive. Adjusting this threshold is also known as threshold tuning. The 50% threshold led to a large number of false positives, so many that they outnumbered the true positives. We applied a threshold tuner (i.e., we adjusted our network so that a much higher result (e.g., 0.9) would be needed for a prediction to be considered a positive) that found the threshold which minimized the false positive rate. This reduced the false positive rate to zero at the individual network level. Threshold tuning was also applied at the ensemble level to the combined score and further optimized to minimize false positives.

The ML algorithm correctly identified 55.6% of COs. In total, 20% of these had a quality flag of 4, 20% had a quality flag of 3, and the remaining 60% had a quality flag of 2. Furthermore, the number of false positives is zero, which suggests all of the objects flagged by our network are COs that fall in the region of outliers shown in Figure 2 and can be appropriately corrected. The final accuracy of the network is 99.86%, outperforming the unbalanced nature of the data (remember the testing and validation data are as unbalanced as the original data). Our results are more accurately reported as a true positive rate of 55.6% and a true negative rate of 100%. Evaluation of the efficacy of ML classification algorithms is typically done using the receiver operator characteristic (ROC) and measuring the area under its curve. We supply ROC curves for both the best constituent model in our network and the ensemble as a whole in Figure 3. The extremely high value of the ROC for our individual algorithms (on average near 0.9 and the best at 0.94) and our ensemble (0.78) indicates that our model performs well and only fails to identify all of the COs in an effort to eliminate false positives. The dramatic reduction in performance from the individual networks to the 0.78 of our ensemble is an artifact of how the ensemble was aggregated. Any model which failed to produce a threshold that would yield 0 false positives was forced to zero. Therefore, all of the thresholds for the ensemble have a false positive rate of zero, artificially reducing the area under the curve of the graph compared to the value had thresholds been used in the individual models that allow for a nonzero false positive rate. Finally, the step-wise nature of the curves is caused by the low number of true positives (a feature of the unbalanced data set) and is not by itself a cause for concern. Figure 3 also shows how much contamination is caused by false positives as the threshold is lowered to achieve a higher true positive rate.

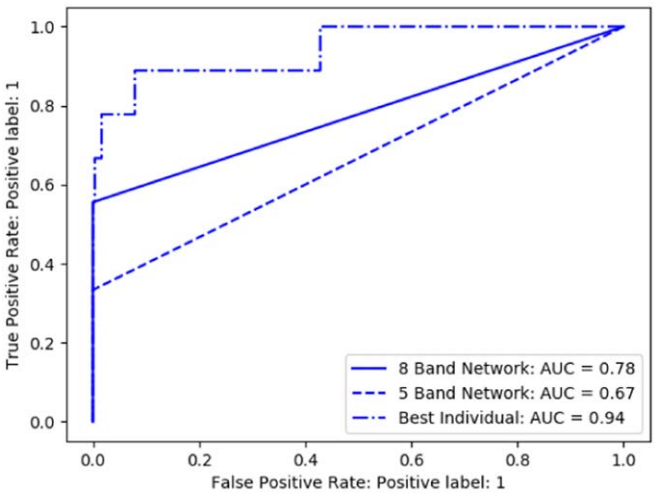

Figure 3. This plot shows the ROC for the eight bandpass network, five bandpass network, and the best individual performing model within the eight bandpass model. Their area under the curve (AUC) scores are 0.78, 0.67, and 0.94, respectively, out of a maximum possible 1.0. The scores of the combined ensembles are generally lower than their individual constituents (e.g., the 0.78 for the eight bandpass vs. its individual 0.94) due to artificial restrictions placed on the ensembles themselves. Because the values coming into the ensemble scoring function have already been restricted to those constituent networks with a false positive rate of 0, only a false positive rate of 0 or 1 is possible. This has artificially reduced the ensembles' ROC AUC scores.

3.3. Training and Validating the Network: Five Bands

Upcoming photo-z challenges (e.g., LSST M. L. Graham et al. 2018) will unfortunately not include the K infrared band, which is crucial for distinguishing between different spectral breaks. Because of this, we repeated our entire analysis utilizing the same training data, network implementation, and architecture to see if our success could be duplicated with fewer bands. Specifically, this new network only utilized the u, g, r, i, and z bands and attempted to classify the same COs and NCOs as above. All treatment of data, the number of networks in the ensemble, the randomization techniques, model hyperparameters, and threshold tuning were identical between the two networks. This process resulted in a network which was able to

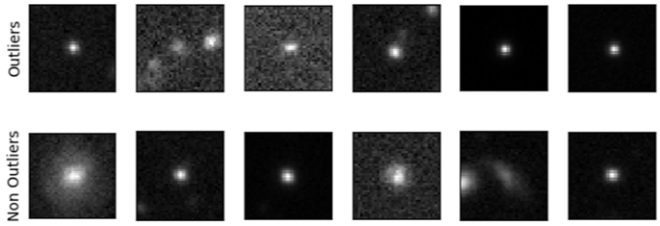

Figure 4. These images are a subsample of the examined postage-stamp images used to check the catalog. There are six outliers and six nonoutliers. These postage stamps are of the I band with a magnitude limit of $I < 25$, and were cut with a half-width of 3″ on each side. These images illustrate the challenge of discerning what might be causing outliers; there appear to be contaminating objects in both COs and NCOs (subfigures 2, 4, and 11 numbered left to right, top to bottom); and both NCOs and COs can also represent a range of compact and distributed objects.

identify 33.3% of COs. Of these, 33.3% had a quality flag of 4, 33.3% had a quality flag of 3, and 33.3% had a quality flag of 2. The ROC curve for this new network is also provided in Figure 3. The performance of this network is worse than the performance of the other network which is unsurprising given the further restrictions on the number of bands. Hereafter, the two networks will be called the eight band network and the five band network.

Our results assume that spectroscopic redshifts are more accurate than their photometric counterparts—as is typical (R. Beck et al. 2016), but we acknowledge this as a limitation of our study in Section 6. We can attempt to correct the identified outliers using the trend line from Figure 2. This trend line is a function of variables photo-z and spec-z of the form photo-$x = m$(spec-z) + b. Our ML algorithm identifies objects that would fall in this region, and we have determined that these objects follow this trend line to a certain degree. Examples of both NCOs and COs are provided in Figure 4. Figure 5 shows the results from our attempt to correct the COs. The corrections and results are discussed further in Section 5.

4. The Cause of Catastrophic Outliers

While this work primarily focuses on identifying COs using ML, we also attempt to discern the cause of COs, which remains unknown. During the preliminary stages of our study, we analyzed both the photometric catalog and spent time analyzing the cutouts of the objects using data from the Hyper Suprime-Cam Subaru Strategic Program. Potential sources of error that could cause discrepancies between the photometric and spectroscopic redshift include the compaction of the object, object morphology, and contamination from nearby sources, particularly if those sources cannot be resolved individually. However, upon visual inspection of a large random sample of the NCOs, and all of the COs at different magnitude cuts, there was no notable difference between the fraction of objects that were faint sources, poorly resolved from nearby contaminants, or varying morphologies and compactness. We have included 12 postage stamps in Figure 4 showing six sample objects of each class.

We further analyzed the available tabular data (ABS Magnitudes) and from these also calculated several colors to evaluate the correlations between magnitudes, colors, and outlier classification. These analyses yielded little fruit in determining the source of the cause of COs. We have included charts showing the relationships between various colors and magnitudes in Figures 6 and 7. These figures demonstrate that the relationship between outliers and nonoutliers is not easily discerned in two dimensions and requires the multidimensional specialty of ML, specifically deep neural networks, which are better at extracting features than more standard regression analyses. We also produced a correlation matrix in Figure 8 to calculate the linear correlations between the variables in our study and outlier status. Unfortunately, there were no strong correlations between variables and the outlier status that could identify the source of COs. Figures 6, 7, and 8 are large and were therefore placed at the end of the text.

Source catalog selection and quality flags are discussed in Section 2, which also has plots detailing the distribution of photo-zs and the selected catastrophic outlier range and details of our sampling techniques. The results of these analyses are summarized in Section 6, and a detailed correlation matrix can be found at the end of the text.

5. Applying the Network

After assessing our networks' accuracy, we applied our networks to the remaining photometric data which do not have spec-zs but do have photometric data for the bandpasses we used to train them. For the eight band network, the data set initially contained 1,669,429 objects, of which 529,934 met all of the criterion in our eight bandpasses and were within the range in photo-z of our training data set $0 \leqslant$ photo-z $\leqslant 6$. Of these, our eight bandpass classifier flagged 83,097 objects as COs after we discarded any of the outliers with a photo-z that were not in the range we defined as our outliers (photo-z [0, 1.5]). This is approximately ≈15.7% of the data which met all of our criteria. However, the ML algorithm misses 45.4% of COs, implying the true fraction of COs could be as high as 28.7% in the most extreme case.

For the five band network, 587,395 objects met all of the criteria for our five bandpasses (u, g, r, i, and z). Of these, 59,544 were flagged as COs by our ML classifier after again discarding any that did not have an original photo-z between [0–1.5]. This is approximately ≈10.1% of the data which met our criteria. However, this ML network misses approximately 66.6% of COs implying the true number of COs could be 30.3% in the most extreme case for these criteria. Both of these results are significantly higher than many other CO predictions based on similar data, motivating further studies to determine what the true fraction is and how to properly account for them.

Directly comparing the objects flagged by the eight band and five band algorithms (83,097 and 59,544 respectively), 44,601 objects were flagged as outliers by both the five band algorithm and the eight band algorithm; 14,472 objects were flagged by the eight bandpass algorithm and flagged by some of the constituent members of the five bandpass algorithm but not enough to be considered an outlier by the five bandpass algorithm as a whole; and 421 objects were flagged by the five bandpass algorithm and flagged by some of the constituent members of the eight bandpass algorithm, but not enough to be considered an outlier by the eight bandpass algorithm as a whole. These numbers are not surprising, the 14,472 objects that were not flagged by the five band network may have been had the other available bands been provided, and the 421 additional outliers found by five band network are likely the product of the five band network considering roughly 60,000 more objects than the eight band network.

Figure 5 presents a set of histograms depicting the original photo-zs and "corrected" photo-zs for both the eight bandpass

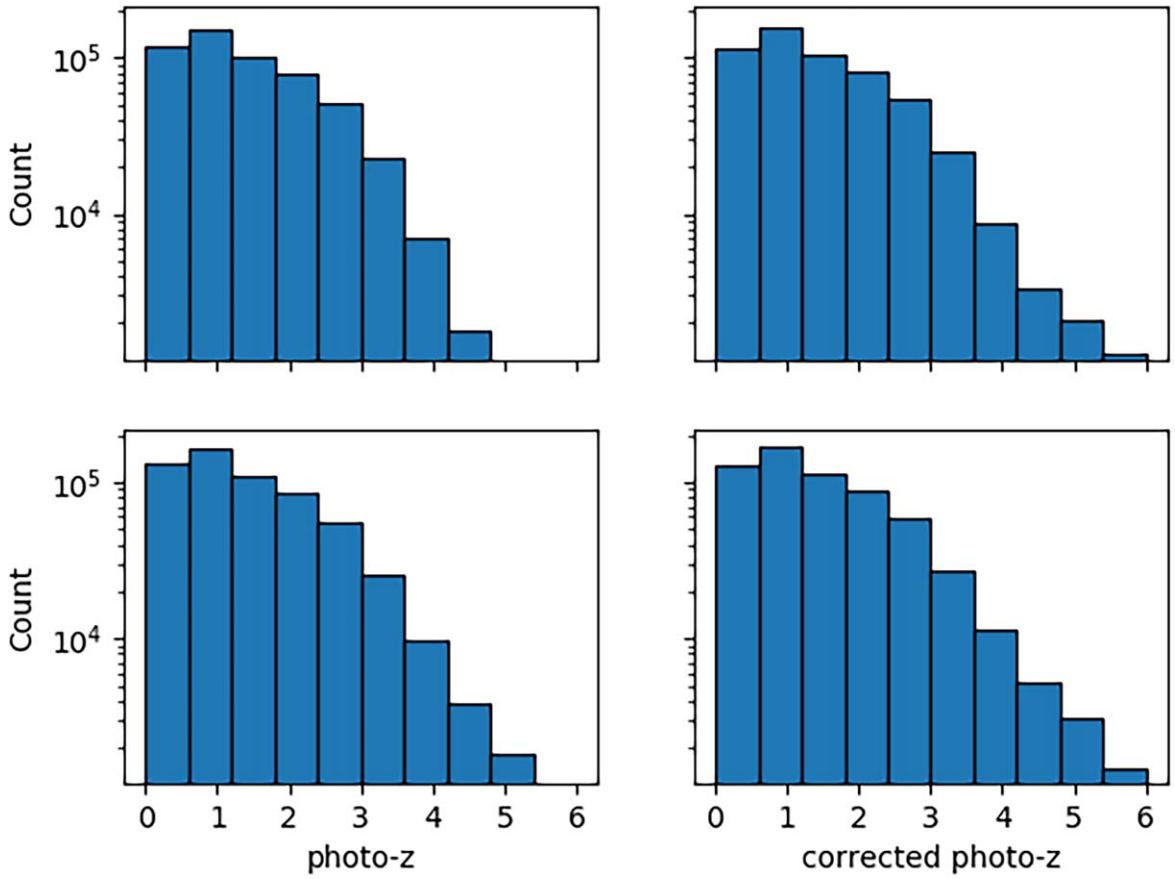

Figure 5. Four histograms comparing the original photometric redshifts from the data not used in training (no available spectroscopic redshifts) on the left with the corrected redshifts, obtained by applying the trend line correction from Figure 2, on the right for objects flagged by the eight bandpass network (top row) and the five bandpass network (bottom row). The slight shift toward higher redshifts indicates a dramatic increase at high redshifts (note the change in the base of the y-axis from left to right). For the eight bandpass network, the number of objects with a photo-$z > 2$ increases from ∼21% to ∼28%, photo-$z > 3$ from ∼5% to ∼10%, photo-$z > 4$ from ∼0.7% to ∼3.5%, and for photo-$z > 5$ from ∼0.11% to ∼1.6%. For the five bandpass network, the number of objects with a photo-$z > 2$ increases from ∼22% to ∼27%, photo-$z > 3$ from ∼6.4% to ∼9.7%, photo-$z > 4$ from ∼1.3% to ∼3.6%, and for photo-$z > 5$ from ∼0.2% to ∼1.8%. For this analysis we only considered objects that, after the correction, fell within the original photo-z range. If all values of photo-z are considered, these ratios become slightly larger.

network (top row) and the five bandpass network (bottom row). The photo-z of the objects flagged by our eight band network as COs can be adjusted using the inverse of the trend line shown in Figure 2—now of the form $\text{spec}-z = \frac{\text{photo}-z - b}{m}$ as a transformation to bring the photo-z from a CO to a value more closely resembling what the spec-z would be. This work does not claim definitive accuracy for these values but it does provide a rough estimate of where a CO's true redshift is, again under the assumption the spectroscopic redshift values are more reliable. We caution readers that this trendline is a toy model and that these values should not be used as photometric redshift. As previously stated, the linear model is not able to obtain highly accurate photo-z measurements but instead can provide a rough picture of how many very high photo-zs are being under-reported.

For the eight bandpass network, the histograms in Figure 5 illustrate an increase in the ratio of objects with a photo-z greater than 2 from 20.8% to 28.0%, an increase in objects with a photo-z greater than 3 from 5.4% to 9.9% , an increase in objects with a photo-z greater than 4 from 0.67% to 3.5%, and an increase in objects with a photo-z greater than 5 from 0.11% to 1.6%. The numbers for the five bandpass histograms are as follows. The ratio of objects with a photo-z greater than 2 increases from 22.3% to 26.4%, greater than 3 increases from 6.4% to 8.8%, greater than 4 increases from 1.3% to 3.6%, and greater than 5 increases from 0.24% to 1.8%.

These numbers utilize only the data from the 529,934 and 587,395 objects that met our selection criteria, and we could only use our toy model on objects that our models flagged. Because our classifiers are not perfect, there may be more objects within these two sets that have low reported photo-zs, increasing these ratios. Conversely, assuming there are no COs in the remaining data, these changes could be reduced by as much as a factor of 3.1 or 2.8 respectively, assuming the distribution of photo-zs remains the same.

These results are a substantial modification to the COSMOS results and are explored further in Section 6.

6. Discussion

6.1. Limitations of Our Study

The first limitation of this study is our assumption that spectroscopic redshifts are consistently more reliable than their photometric counterparts. We aimed to address this by only using the spectroscopic data that had higher quality flags, but the fact remains that some of these spectroscopic redshifts may be incorrect, and therefore our results overstated. The results of our testing data indicate half of the outliers may correspond to objects that would have a spectroscopic quality flag of 2 (the

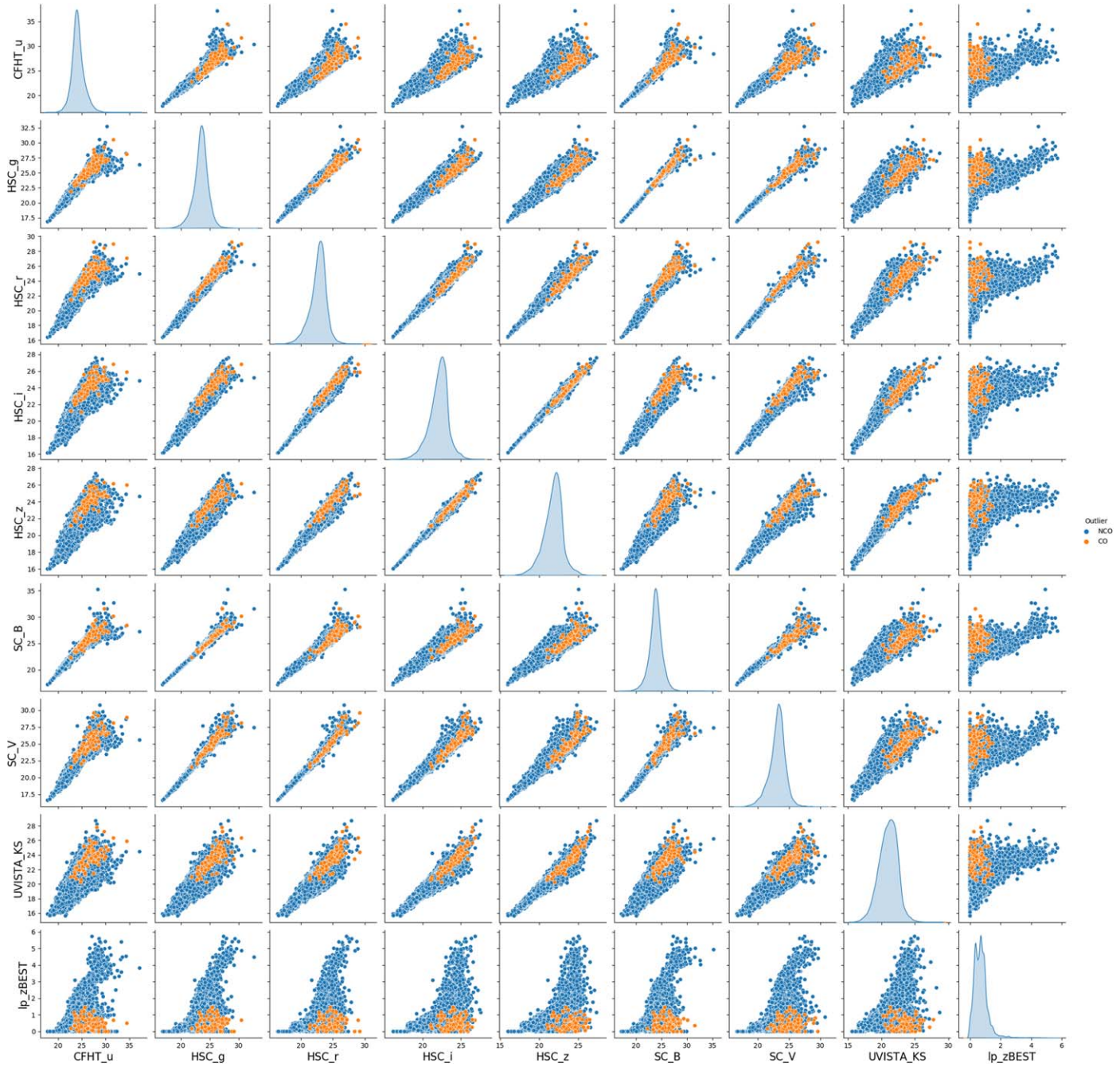

Figure 6. A series of scatterplots and kernel density estimate (KDE) plots pairing each of the filter magnitudes together. The plots along the diagonal contain KDE plots for each respective filter (and the photometric redshift). The scatterplots contain data corresponding to the x and y labels of each row or column. The upper triangle is simply a reflection of the lower triangle. These charts show almost no meaningful distinction between COs and NCOs in the magnitudes alone.

other half having a quality flag of 4). Although there is no way to filter these from our predictions, we can conservatively estimate that up to half of our identified COs in the prediction data may ultimately have less reliable spectroscopic redshifts.

Another limitation of our work is the very small number of true positives in our data set. The number of NCOs is relatively large, enabling the ML to clearly label it in parameter space. However, the small number of COs makes the challenge of carefully identifying the true boundary of the CO space even harder. Because of this, it is possible that the ML is predicting false positives in the photometric data, even when the testing data returns a false positive rate of zero. This would also result in an overstatement of our results.

Additionally, our study focused on data with a detection in all of the available bandpasses. However, this eliminated a substantial fraction of the total data (≈68% for the eight bandpass network and ≈65% for the five bandpass network). Assuming the most conservative case (there are no COs in this data), this would reduce the number of COs detected to roughly 4.9% for the eight bandpass network and 3.5% for the five bandpass network or 8.9% and 10.7% when accounting for the missed objects, respectively. If these are further combined with our first limitation, then we can conservatively estimate that the lowest rates of COs from our two results are ≈3.6% and ≈7.1% (only allowing for quality flags of 3 or 4), i.e., a 60% reduction from 8.9% and a 33.3% reduction from 10.7%, for

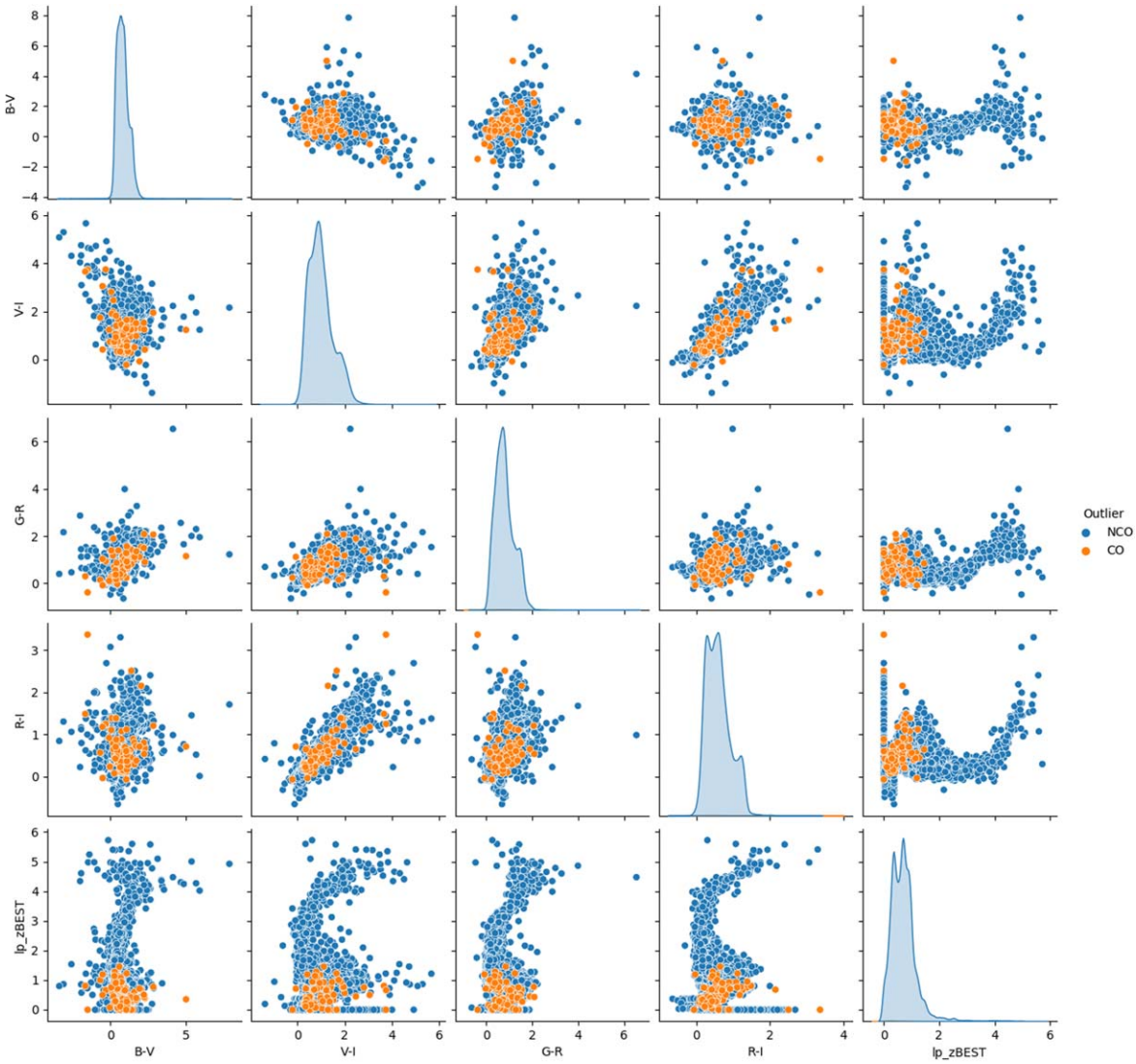

Figure 7. The same as Figure 6 but with four common colors instead of the magnitudes.

the eight band and five band networks, respectively. However, considering that the other half of the objects have fewer detected bandpasses, and given the typical assumption that spectroscopic redshifts are more reliable than photometric redshifts, this most conservative estimate would seem unlikely.

Finally, it is important to note that while our results indicate an increase in the number of objects recorded at redshift greater than 2, other work focusing on the inverse of our area of interest (high photo-z but low spec-z) will almost certainly decrease the number of objects recorded at redshift greater than 2. Since we only focused on one of these regions, the difference between the increase and decrease is beyond the scope of this work.

6.2. Comparison with Other Works

Our work suggests that the fraction of objects within our combined data set which may be COs is at least 3.6% in the most conservative case, with a significantly larger fraction than estimated by other works (e.g., 0.5% C. Laigle et al. 2016). This has potentially large ramifications for the confidence placed in photometric redshifts, particularly at the high end of the redshift regime. While our work does not correctly identify the same fraction of outliers (we note these other works use different definitions of COs) as other ML methods (e.g., J. Singal et al. 2022) at roughly 80% for some small number redshifts, more typically near 50% based on their Figure 2 compared to our 55%, our work achieves a false positive rate of 0. However, our work does match their 80% true positive rate at a false positive threshold of only 10% for our best network (and comparable values of the true positive rate and false positive rates in the other networks). In other works, the number of COs correctly identified, particularly in the area of interest for our study ($\approx z$-spec > 2), is a factor of 10 or greater less than the number of nonoutliers, making potential spectroscopic follow-up to confirm these results expensive with many false positives. Other works (e.g., T. Dahlen et al. 2008; V. E. Margoniner & D. M. Wittman 2008; S. J. Schmidt & P. Thorman 2013; M. Wyatt & J. Singal 2021), whose results are thoroughly summarized in M. Wyatt & J. Singal (2021), use probability information to flag COs, but they also incorrectly flag a large number of NCOs to a degree that is likely similarly prohibitive to follow-up observations targeting COs. Again, we remind the reader that these other works use different definitions of COs. Our definition is the most restrictive, focusing exclusively on the high spec-z, low photo-z range.

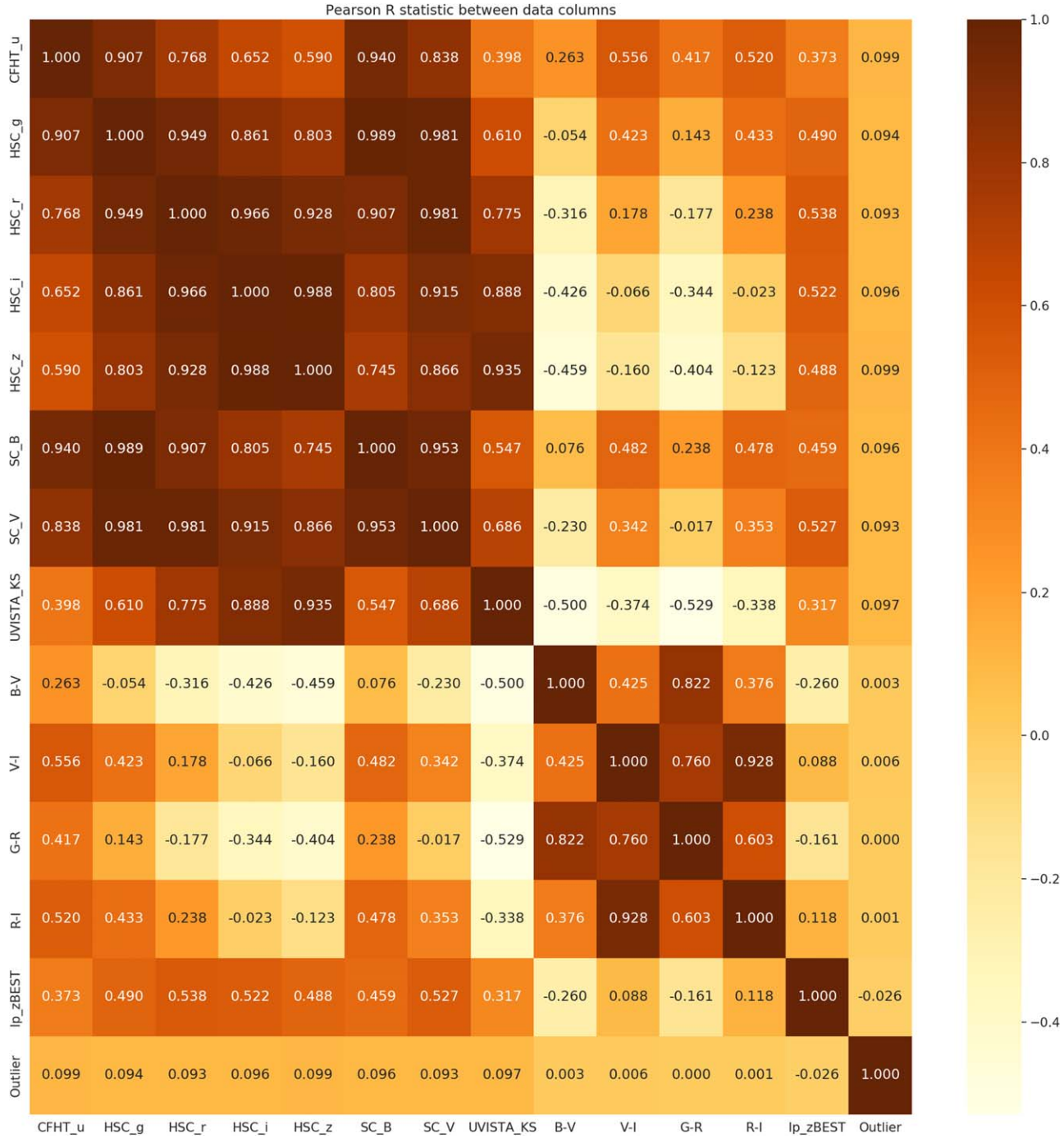

Figure 8. A correlation heatmap with data columns on the *x* and *y* axes and the Pearson R statistic labeled at each of the grid locations. The row and column labels correspond to the *B*, *V*, *u*, *r*, *i*, *z*, and *Ks* magnitudes, their associated errors, and the photo-*z*, and the upper and lower 68% photo-*z* confidence intervals. While there are strong correlations between the various magnitude columns, and there are strong correlations between the photo-*z* and the 68% confidence interval bounds, there are no strong correlations between any of the columns with our target variable (CO). Without strong correlations between the data columns, more advanced techniques such as ML are needed to accurately identify COs.

7. Conclusions

The primary result of this work is the construction of an ensemble of neural networks that accurately identifies COs with a true positive rate of 55.6% for eight bandpasses and 33.3% for five bandpasses, both with a false positive rate of 0%. Additionally, these ensembles identified a list of 83,097 and 59,544 targets, respectively, which may also be previously unidentified COs within the COSMOS2020 data set. Analysis of the networks suggests the most conservative estimate for the number of COs fitting our criteria is ≈3.6%. Finally, when these are corrected with a toy regression model, it shifts the fraction of objects with photometric redshifts greater than 5 from 0.11% to 1.6% for the eight band network and 0.24% to 1.8% for the five band network.

Developing methods to correct COs within large photometric catalogs is crucial. Since there is not enough telescope time to provide spectroscopically determine redshifts, it is necessary to

be able to develop methods that can provide both accurate photo-zs and also other methods that can independently determine when these photo-z algorithms are incorrect. Research conducted by J. Singal et al. (2022) demonstrates the feasibility of creating neural networks and other ML algorithms capable of accurately identifying a substantial fraction of COs within photometric redshift catalogs (note the difference in the definition of COs in their work). Our work takes this a step further by both developing a ML algorithm that correctly identifies a fraction of COs, while also doing it with few to no false positives. Because of the large imbalance between the NCOs and COs, a false positive rate of 0 or of an order where the false positives do not outnumber the true positives is helpful for identifying potential follow-up targets to validate the results of our classifier. Conducting additional spectroscopic and photometric follow-up observations on these COs will provide researchers with more samples to develop photo-z templates and/or ML algorithms that can accurately identify and correct COs in large galactic surveys.

Implementing data-driven corrections, as introduced in this work and others, can significantly improve photometric redshift estimates, thereby enhancing their reliability. Improving the reliability of photometric redshifts also has consequences for cosmology, where photometric redshifts have been used to draw conclusions about the large-scale structure of the universe. Furthermore, our work only applies a simple correction to the identified COs (fitting to a trend line). As more of these types of objects are identified, more sophisticated methods could be developed (e.g., a custom SED template set) to improve these corrections further.

Future work on the COSMOS data set could include the inverse of our designated outliers, (i.e., the objects with high reported phot-zs and low spec-zs). Another possible direction would be the incorporation of images of these objects. The reduction of the data to magnitudes could easily exclude useful information that may help the ML identify COs (e.g., nearby contaminants, compactness, etc).

Beyond the COSMOS data set, future work could expand to other large data sets such as the North Ecliptic Plane (which has data in similar band passes) and other large photometric sky surveys. Furthermore, the CO problem is one example of the "rare object detection problem" and a case study in unbalanced data sets. As modern instruments continue to collect thousands of images on millions of objects, there is a rapidly growing need for tools to pick out rare objects (e.g., metal-poor and extremely metal-poor galaxies) within these data sets (e.g., A. Soto et al. 2005). While further work does need to be done to address the unbalanced data set problem, our work shows it is possible to apply ML methods to find rare objects within these large astronomical data sets with a high degree of accuracy.

Acknowledgments

We gratefully acknowledge support for this research from NSF grant AST-1716093 (E.M.H., M.D.). We thank Amy Barger, Anthony Taylor, and Peter Sadowski for their helpful comments and suggestions.

Software: `NumPy` version 1.22.3 (C. R. Harris et al. 2020), `Pandas` version 1.4.3 (W. McKinney 2010; The pandas development team 2020), `Matplotlib` version 3.5.1 (J. D. Hunter 2007), `Seaborn` version 0.12.0, (M. L. Waskom 2021), `TensorFlow` version 2.4.1 (M. Abadi et al. 2015), Keras version 2.8.0 (F. Chollet et al. 2015), `SciKit-Learn` version 1.0.2 (F. Pedregosa et al. 2011), `imbalanced-learn` version 0.9.0 (G. Lemaître et al. 2017).

ORCID iDs

Mitchell T. Dennis https://orcid.org/0000-0001-9066-0552
Esther M. Hu https://orcid.org/0009-0008-7427-4617
Lennox L. Cowie https://orcid.org/0000-0002-6319-1575

References

Abadi, M., Agarwal, A., Barham, P., et al. 2015, TensorFlow: Large-scale Machine Learning on Heterogeneous Systems, https://www.tensorflow.org/
Arnouts, S., Cristiani, S., Moscardini, L., et al. 1999, MNRAS, 310, 540
Beck, R., Dobos, L., Budavári, T., Szalay, A. S., & Csabai, I. 2016, MNRAS, 460, 1371
Benítez, N. 2000, ApJ, 536, 571
Bernstein, G., & Huterer, D. 2010, MNRAS, 401, 1399
Blanton, M. R., Bershady, M. A., Abolfathi, B., et al. 2017, AJ, 154, 28
Bolzonella, M., Miralles, J. M., & Pelló, R. 2000, A&A, 363, 476
Brammer, G. B., van Dokkum, P. G., & Coppi, P. 2008, ApJ, 686, 1503
Brescia, M., Cavuoti, S., Razim, O., et al. 2021, FrASS, 8, 70
Capak, P., Aussel, H., Ajiki, M., et al. 2007, ApJS, 172, 99
Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. 2011, arXiv:1106.1813
Chollet, F., et al. 2015, Keras, https://keras.io
Cooper, O. R., Casey, C. M., Akins, H. B., et al. 2024, ApJ, 970, 50
Dahlen, T., Mobasher, B., Jouvel, S., et al. 2008, AJ, 136, 1361
Dickinson, M., Giavalisco, M. & GOODS Team 2003, in The Mass of Galaxies at Low and High Redshift, ed. R. Bender & A. Renzini (Berlin: Springer), 324
Dobos, L., Csabai, I., Yip, C.-W., et al. 2012, MNRAS, 420, 1217
Foo, F. F., Yang, A., & Chan, A. H. 2014, in Proc. Conf. in Honour of the 90th Birthday of Freeman Dyson, ed. K. K. Phua et al. (Singapore: World Scientific), 463
Graham, M. L., Connolly, A. J., Ivezić, Ž., et al. 2018, AJ, 155, 1
Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Natur, 585, 357
Hasinger, G., Capak, P., Salvato, M., et al. 2018, ApJ, 858, 77
Hunter, J. D. 2007, CSE, 9, 90
Ilbert, O., Arnouts, S., McCracken, H. J., et al. 2006, A&A, 457, 841
Ilbert, O., Capak, P., Salvato, M., et al. 2008, ApJ, 690, 1236
Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111
Jafariyazani, M., Masters, D., Faisst, A. L., Teplitz, H. I., & Ilbert, O. 2024, ApJ, 967, 60
Jones, E., & Singal, J. 2017, A&A, 600, A113
Jones, E., & Singal, J. 2020, PASP, 132, 024501
Kingma, D. P., & Ba, J. 2014, arXiv:1412.6980
Laigle, C., McCracken, H. J., Ilbert, O., et al. 2016, ApJS, 224, 24
Lemaître, G., Nogueira, F., & Aridas, C. K. 2017, JMLR, 18, 1
Lilly, S. J., Le Fèvre, O., Renzini, A., et al. 2007, ApJS, 172, 70
Margoniner, V. E., & Wittman, D. M. 2008, ApJ, 679, 31
McKinney, W. 2010, in Proc. 9th Python in Science Conf., ed. S. van der Walt & J. Millman (Austin, TX: SciPy), 56
Momtaz, A., Salimi, M. H., & Shakeri, S. 2022, arXiv:2201.04391
Pasquet, J., Bertin, E., Treyer, M., Arnouts, S., & Fouchez, D. 2019, A&A, 621, A26
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, JMLR, 12, 2825
Pozzetti, L., Bolzonella, M., Zucca, E., et al. 2010, A&A, 523, A13
Rudnick, G., Franx, M., Rix, H.-W., et al. 2001, AJ, 122, 2205
Schmidt, S. J., & Thorman, P. 2013, MNRAS, 431, 2766
Sillassen, N. B., Jin, S., Magdis, G. E., et al. 2024, A&A, 690, A55
Singal, J., Silverman, G., Jones, E., et al. 2022, ApJ, 928, 6
Soto, A., Cansado, A., & Zavala, F. 2005, in ASP Conf. Ser. 347, Astronomical Data Analysis Software and Systems XIV, ed. P. Shopbell, M. Britton, & R. Ebert (San Francisco, CA: ASP), 66
Sui, Y., Zhang, X., Huan, J., & Hong, H. 2019, Proc. SPIE, 11198, 1119813
Taylor, A. J., Barger, A. J., Cowie, L. L., et al. 2023, ApJ, 266, 24
The pandas development team 2020, pandas-dev/pandas: Pandas, Latest, Zenodo, doi:10.5281/zenodo.3509134
Thorp, S., Alsing, J., Peiris, H. V., et al. 2024, ApJ, 975, 145
Waskom, M. L. 2021, JOSS, 6, 3021
Weaver, J. R., Kauffmann, O. B., Ilbert, O., et al. 2022, ApJS, 258, 11
Wyatt, M., & Singal, J. 2021, PASP, 133, 044504