Using Multivariate Imputation by Chained Equations to Predict Redshifts of Active Galactic Nuclei

Spencer James Gibson[1], Aditya Narendra[2], Maria Giovanna Dainotti[3,4]*, Malgorzata Bogdan[5,6], Agnieszka Pollo[7,8], Artem Poliszczuk[8], Enrico Rinaldi[9,10,11] and Ioannis Liodakis[12]

[1]Carnegie Mellon University, Pittsburgh, PA, United States, [2]Astronomical Observatory of Jagiellonian University, Kraków, Poland, [3]National Astronomical Observatory of Japan, Mitaka, Japan, [4]Space Science Institute, Boulder, CO, United States, [5]Department of Mathematics, University of Wroclaw, Wrocław, Poland, [6]Department of Statistics, Lund University, Lund, Sweden, [7]Astronomical Observatory of Jagiellonian University, Krakow, Poland, [8]National Centre for Nuclear Research, Warsaw, Poland, [9]Physics Department, University of Michigan, Ann Arbor, MI, United States, [10]Theoretical Quantum Physics Laboratory, RIKEN, Wako, Japan, [11]Interdisciplinary Theoretical and Mathematical Science Program, RIKEN (iTHEMS), Wako, Japan, [12]Finnish Centre for Astronomy with ESO (FINCA), University of Turku, Turku, Finland

Redshift measurement of active galactic nuclei (AGNs) remains a time-consuming and challenging task, as it requires follow up spectroscopic observations and detailed analysis. Hence, there exists an urgent requirement for alternative redshift estimation techniques. The use of machine learning (ML) for this purpose has been growing over the last few years, primarily due to the availability of large-scale galactic surveys. However, due to observational errors, a significant fraction of these data sets often have missing entries, rendering that fraction unusable for ML regression applications. In this study, we demonstrate the performance of an imputation technique called Multivariate Imputation by Chained Equations (MICE), which rectifies the issue of missing data entries by imputing them using the available information in the catalog. We use the Fermi-LAT Fourth Data Release Catalog (4LAC) and impute 24% of the catalog. Subsequently, we follow the methodology described in Dainotti et al. (ApJ, 2021, 920, 118) and create an ML model for estimating the redshift of 4LAC AGNs. We present results which highlight positive impact of MICE imputation technique on the machine learning models performance and obtained redshift estimation accuracy.

Keywords: redshift, AGNs, BLLs, FSRQs, FERMI 4LAC, machine learning regressors, imputation, MICE

1 INTRODUCTION

Spectroscopic redshift measurement of Active Galactic Nuclei (AGNs) is a highly time-consuming operation and is a strong limiting factor for a large-scale extragalactic surveys. Hence, there is a pressing requirement for alternative redshift estimation techniques that provide reasonably good results Salvato et al. (2019). In current cosmological studies, such alternative redshift estimates, referred to as photometric redshifts, play a key role in our understanding of the Extragalactic Background Light (EBL) origins Wakely and Horan (2008)1, magnetic field structure in the intergalactic medium Marcotulli et al. (2020); Venters and Pavlidou (2013); Fermi-LAT Collaboration et al. (2018) and help in determining the bounds on various cosmological parameters Domínguez et al. (2019); Petrosian (1976); Singal et al. (2013b), Singal et al. (2012),

Singal et al. (2014); Singal (2015); Singal et al. (2013a); Chiang et al. (1995); Ackermann et al. (2015); Singal et al. (2013b); Ackermann et al. (2012).

One technique that has gained significant momentum is the use of machine learning (ML) to determine the photometric redshift of AGNs Brescia et al. (2013), Brescia et al. (2019); Dainotti et al. (2021); Nakoneczny et al. (2019); Jones and Singal (2017); Cavuoti et al. (2014); Fotopoulou and Paltani (2018); Logan and Fotopoulou (2020); Yang et al. (2017); Zhang et al. (2019); Curran (2020); Nakoneczny et al. (2019); Pasquet-Itam and Pasquet (2018); Jones and Singal (2017). Large AGN data sets derived from all-sky surveys like the Wide-field Infrared Survey Explorer (WISE) Brescia et al. (2019); Ilbert et al. (2008); Hildebrandt et al. (2010); Brescia et al. (2013); Wright et al. (2010); D'Isanto and Polsterer (2018) and Sloan Digital Sky Survey (SDSS) Aihara et al. (2011) have played a significant role in the proliferation of ML approaches. However, the quality of the results from an ML approach depends significantly on the size and quality of the training data: the data on which the ML models learn the underlying relationship to predict the redshift. Unfortunately, almost all of these large data sets suffer from the issue of missing entries, which can lead to a considerable portion of the data being discarded.

This is especially problematic in catalogs of smaller size, such as in the case of gamma-ray loud AGNs.

Using the Fermi Fourth Data Release Catalog's (4LAC) gamma-ray loud AGNs Ajello et al. (2020); Abdollahi et al. (2020), Dainotti et al. (2021) demonstrated that ML methods lead to promising results, with a 71% correlation between the predicted and observed redshifts. However, in that study, the training set consists of only 730 AGNs, and a majority of the data (50%) are discarded due to missing entries. More specifically, we have several reasons why the sources are missing also in relation to the variables we consider. Regarding the missing values of the Gaia magnitudes: this could be either because the sources are too faint and thus they undergo the so called Malmquist bias effect (only the brightest sources are visible at high-z) or the coordinates are not accurate enough and the cross-matching is failing to produce a counterpart (the latter is not that likely, the former is much more likely).

Regarding the variables observed in γ-rays: here the source is detected, but it is faint in gamma-rays and again we have the Malmquist bias effect in relation to the detector threshold of Fermi-LAT and/or it does not appear variable and/or the spectral fitting fails to produce values, hence the missing values.

Regarding the multi-wavelength estimates (ν, νf_ν): these depend on the availability of multi-wavelength data from radio to X-rays. If sufficient data exists then a value can be estimated, so the missing values are most likely sources that have not been observed by telescopes. In other words, this does not mean that the sources are necessarily faint, they could be bright, but just no telescope performed follow-up observations.

There is also the possibility to explain the missing values because of the relativistic effects that dominate blazar emission. The relativistic effects, quantified by a parameter called the Doppler factor, boost the observed flux across all frequencies, but also shorten the timescales making sources appear more variable. It has been shown that sources detected in γ-rays have higher Doppler factors and are more variable Liodakis et al. (2017), Liodakis et al. (2018). This would suggest that sources observed more off-axis, i.e., lower Doppler factor, would have a lower γ-ray flux and appear less variable. Therefore introduce more missing values as we have discussed above.

In this study, we address this issue of missing entries using an imputation technique called Multivariate Imputation by Chained Equations (MICE) Van Buuren and Groothuis-Oudshoorn (2011). This technique was also recently used by Luken et al. (2021) for redshift estimation of Radio-loud AGNs.

Luken et al. (2021) test multiple imputation techniques, MICE included, to determine the best tool for reliably imputing missing values. Their study considers the redshift estimation of radio-loud galaxies present in the Australia Telescope Large Area Survey (ATLAS). However, in contrast to our approach where we impute actual missing information in the catalog, they manually set specific percentages of their data as missing and test how effective various imputation techniques are. Their results demonstrate distinctly that MICE is the best imputation technique, leading to the least root mean square error (RMSE) and outlier percentages for the regression algorithms they have tested.

In our study, we are using the updated 4LAC catalog, and using MICE imputations to fill in missing entries, we achieve a training data set which is 98% larger than the one used in Dainotti et al. (2021). We achieve results on this more extensive training set that are comparable to Dainotti et al. (2021) while attaining higher correlations. Furthermore, we are using additional ML algorithms in the SuperLearner ensemble technique, as compared to Dainotti et al. (2021).

Section 2 discusses the specifics of the extended 4LAC data set: how we create the training set, which predictors are used and which outliers are removed. In **Section 3** we discuss the MICE imputation technique, the SuperLearner ensemble with a brief description of the six algorithms used in this analysis, followed by the different feature engineering techniques implemented. Finally, we present the results in **Section 4**, followed by the discussion and conclusions in **Section 5**.

2 SAMPLE

This study uses the Fermi Fourth Data Release Catalog (4LAC), containing 3,511 gamma-ray loud AGNs, 1764 of which have a measured spectroscopic redshift. Two categories of AGNs dominate the 4LAC catalog, BL Lacertae (BLL) objects and Flat Spectrum Radio Quasars (FSRQ). To keep the analysis consistent with Dainotti et al. (2021), we remove all the non-BLL and non-FSRQ AGNs.

These AGNs have 13 measured properties in the 4LAC catalog; however, we only use 11 and a categorical variable that distinguishes BLLs and FSRQs. The two omitted properties in the analysis are *Highest_Energy* and *Fractional_Variability* because 42.5% of the entries are missing, and there is insufficient information to impute them reliably. We consider imputation of predictors which have

FIGURE 1 | Redshift distribution of the training set.
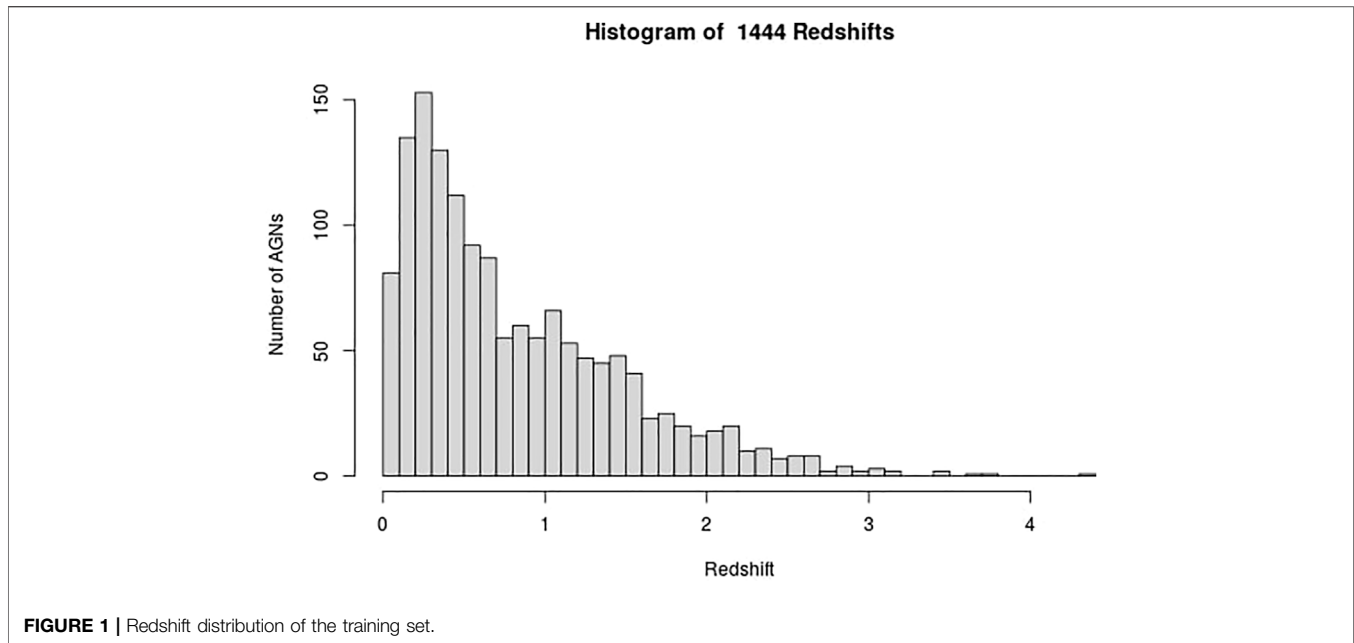
missing entries in less than 18% of the data. The remaining 11 properties and the categorical variables are *Gaia_G_Magnitude*, *Variability_Index*, *Flux*, *Energy_Flux*, *PL_Index*, *νfν*, *LP_Index*, *Significance*, *Pivot_Energy*, *ν*, and *LP_β* and *LabelNo*, serve as the predictors for the redshift in the machine learning models and are defined in Dainotti et al. (2021) and Ajello et al. (2020). However, some of these properties are not used as they appear in the 4LAC, since they span several orders of magnitude. The properties *Flux*, *Energy_Flux*, *Significance*, *Variability_Index*, *ν*, *νfν*, and *Pivot_Energy* are used in their base-10 logarithmic form. In the categorical variable *LabelNo* we assign the values 2 and 3 to BLLs and FSRQs, respectively. We are not training the ML models to predict the redshift directly. Instead, we train the models to predict $1/(z + 1)$, where z is the redshift. Such a transformation of the target variable is crucial as it helps improve the model's performance. In addition, $1/(z + 1)$ is known as the scale factor and has a more substantial cosmological significance than redshift itself. We remove AGNs with an *LP_β* < 0.7, *LP_Index* > 1, and *LogFlux* > -10.5, as they are outliers of their respective distributions. These steps lead us to a final data sample of 1897 AGNs, out of which 1,444 AGNs have a measured redshift (see **Figure 1**). These AGNs form the training sample, while the remaining 453 AGNs, which do not have a measured redshift, form the generalization sample.

3 METHODOLOGY

Here we present the various techniques implemented in the study, definitions of the statistical metrics used, and a comprehensive step-by-step description of our procedure to obtain the results. We use the following metrics to measure the performance of our ML model:

- Bias: Mean of the difference between the observed and predicted values.
- σ_{NMAD}: Normalized median absolute deviation between the predicted and observed measurements.
- r: Pearson correlation coefficient between the predicted and observed measurements.
- Root Mean Square Error (RMSE) between the predicted and observed redshift
- Standard Deviation σ between the predicted and observed redshift

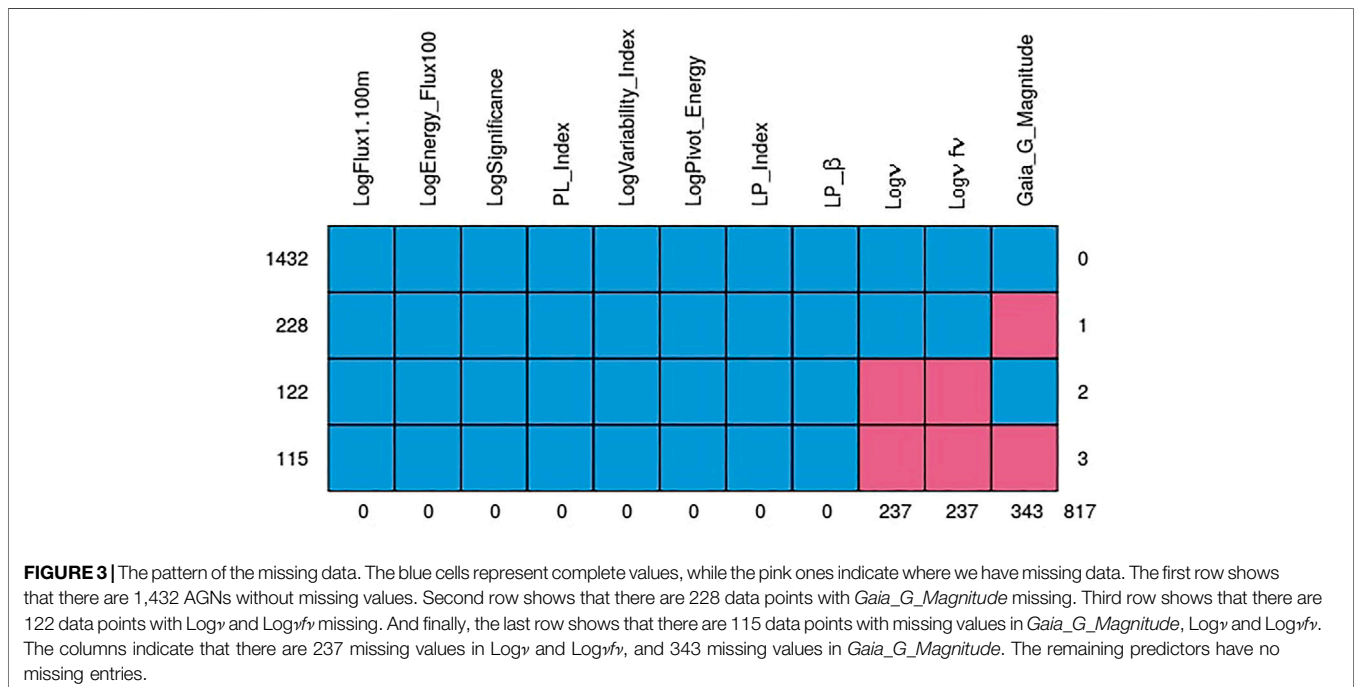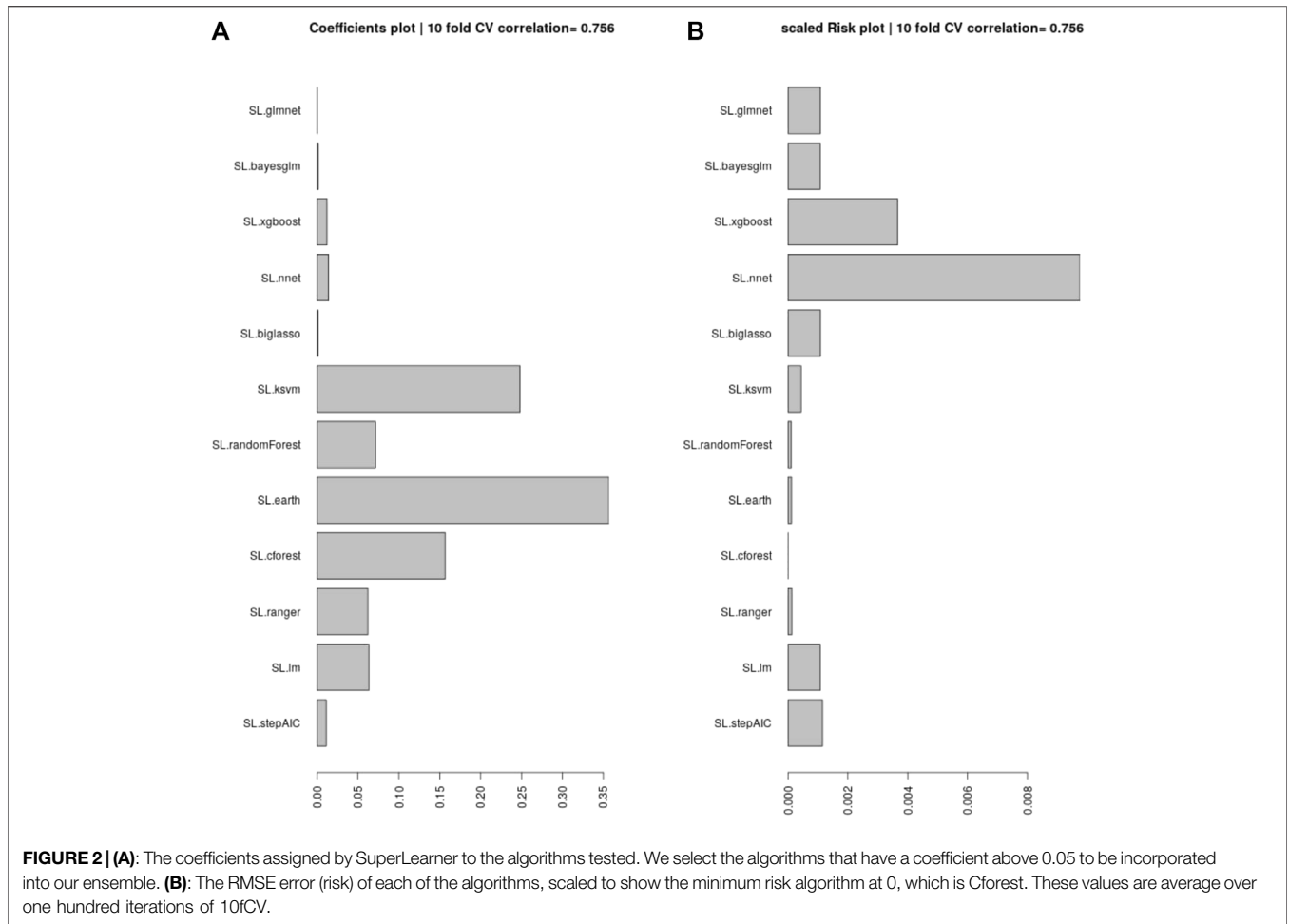We present these metrics for both Δz_{norm} and Δz, which are defined as:
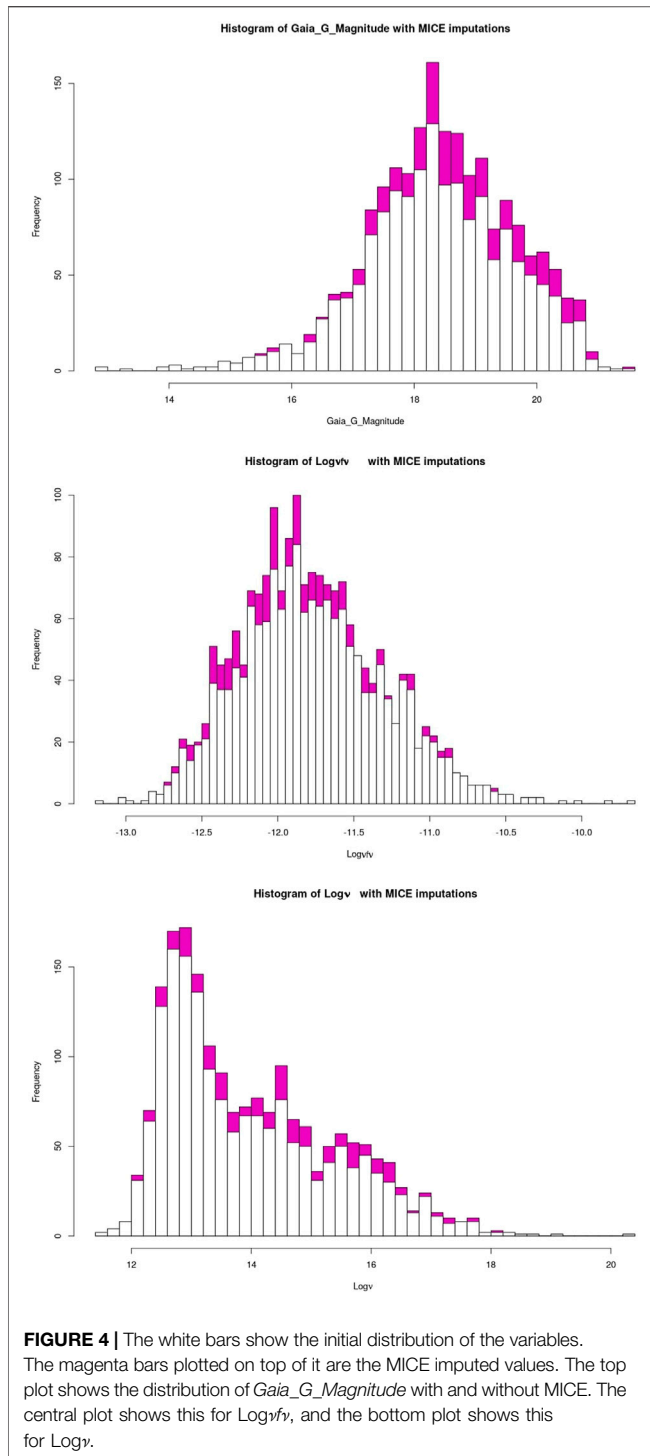
$$\Delta z = z_{observed} - z_{predicted} \tag{1}$$

$$\Delta z_{norm} = \frac{\Delta z}{1 + z_{observed}} \tag{2}$$

We also quote the catastrophic outlier percentage, defined as the percentage of predictions that lies beyond the 2σ error. The metrics presented in this study are the same as in Dainotti et al. (2021), allowing for easy comparison.

3.1 Procedure

Here we provide a walk-through of how the final results are obtained. First, we remove all the non-BLL and non-FSRQ AGNs from the 4LAC data set, in addition to outliers, and end up with 1897 AGNs for the total set. Then, we impute the missing entries using MICE (see **Section 3.2**). Having obtained a complete data set, we split it into the training and the generalization sets, depending on whether the AGNs have or do not have a measured redshift value. We aim to train an ensemble model that is the least complex and best suited to the data at hand. For this purpose, we need to test many different algorithms with ten-

FIGURE 2 | (A): The coefficients assigned by SuperLearner to the algorithms tested. We select the algorithms that have a coefficient above 0.05 to be incorporated into our ensemble. **(B)**: The RMSE error (risk) of each of the algorithms, scaled to show the minimum risk algorithm at 0, which is Cforest. These values are average over one hundred iterations of 10fCV.

FIGURE 3 | The pattern of the missing data. The blue cells represent complete values, while the pink ones indicate where we have missing data. The first row shows that there are 1,432 AGNs without missing values. Second row shows that there are 228 data points with *Gaia_G_Magnitude* missing. Third row shows that there are 122 data points with Logν and Log$\nu f\nu$ missing. And finally, the last row shows that there are 115 data points with missing values in *Gaia_G_Magnitude*, Logν and Log$\nu f\nu$. The columns indicate that there are 237 missing values in Logν and Log$\nu f\nu$, and 343 missing values in *Gaia_G_Magnitude*. The remaining predictors have no missing entries.

FIGURE 4 | The white bars show the initial distribution of the variables. The magenta bars plotted on top of it are the MICE imputed values. The top plot shows the distribution of *Gaia_G_Magnitude* with and without MICE. The central plot shows this for Log$\nu f\nu$, and the bottom plot shows this for Logν.

fold cross-validation (10fCV). Cross-validation is a resampling procedure that uses different portions of the data, in this case 10, to train and test a model, and find out which algorithm performs the best in terms of the previously defined metrics. However, since there is inherent randomness in how the folds are created during 10fCV, we perform 10fCV one hundred times and average the results to derandomize and stabilize them. This repeated

k-fold cross-validation technique is standard in evaluating ML models. In each of the one hundred iterations of 10fCV, we train a SuperLearner model (see **Section 3.3**) on the training set using the twelve algorithms shown in **Figure 2**. Finally, averaging over the one hundred iterations, we obtained the coefficients and risk measurements associated with each SuperLearner ensemble model, as well as the individual algorithms. Following the previous step, we pick six algorithms that have coefficients greater than 0.05 (see **Section 3.4** for information about these algorithms).

With the six best ML algorithms, we create an ensemble with SuperLearner and perform the 10fCV one hundred times once more. The final cross-validated results are again an average of these one hundred iterations.

Next, we proceed to show the results obtained without the repeated cross-validation procedure. For this, we simply select a fixed validation set by choosing the last 111 AGNs from the 1,444 AGNs of the previously used training data. Now, with the new training set of 1,333 AGNs, we train a SuperLearner model, with the algorithms being the same as in the cross-validation step, and we predict the redshift of the validation set. We then calculate the same statistical metrics for these results as we did for the cross-validated results. The results on this fixed validation set provide a representative of the performance of the SuperLearner model, which we have explored in more details (and in a more computationally expensive way) during the repeated cross-validation procedure.

3.2 Multivariate Imputation by Chained Equations

Multivariate Imputation by Chained Equations (MICE) is a method for imputing missing values for multivariate data Van Buuren and Groothuis-Oudshoorn (2011); Luken et al. (2021). The multivariate in MICE highlights its use of multiple variables to impute missing values. The MICE algorithm works under the assumption that the data are missing at random (MAR). MAR was first detailed in the paper Rubin (1976). It implies that errors in the system or with users cause the missing entries and not intrinsic features of the object being measured. Furthermore, MAR implies the possibility that the missing entries can be inferred by the other variables present in the data Schafer and Graham (2002). Indeed, this is a strong assumption, and it is our first step to deal with missing data. However, we know that selection biases play an important role for the flux detection. Although this problem is mitigated for the gamma-ray sources, for the G-band magnitude, one can argue that, e.g., BL Lacs are systematically fainter than FSRQs and below the Gaia limiting magnitude. A more in-depth analysis to take this problem into account is worthwhile, but this is beyond the scope of the current paper.

With this assumption, MICE attempts to fill in the absent entries using the complete variables in the data set iteratively. We impute the missing variables 20 times with each iteration of MICE consisting of multiple steps. General practice is to perform the imputation ten times as in Luken et al. (2021) and Van

TABLE 1 | Composition of the training and generalization sets, and Redshift properties on the training set.

Type	Training set	Generalization set	Redshift median	Redshift minimum	Redshift maximum
BLLs	721	450	0.336	3.7×10^{-5}	2.82
FSRQ	723	3	1.12	0.097	4.313
Total	1,444	453	0.628	3.7×10^{-5}	4.313

Buuren and Groothuis-Oudshoorn (2011), but we perform it twenty times to stabilize the imputation.

Here, we use the method "midastouch"—a predictive mean matching (PMM) method Little and Rubin (2019). It works by initializing a feature's missing entries with its mean and then estimating them by training a model using the rest of the complete data. For each prediction, a probability is assigned based on its distance from the value imputed for the desired entry. The missing entry is imputed by randomly drawing from the observed values of the respective predictor, weighted according to the probability defined previously.

The process is repeated for each missing entry until all have been refitted. This new complete table is used as a basis for the next iteration of MICE, where the same process is repeated until the sequence of table converges or a set number of iterations is achieved.

3.3 SuperLearner

SuperLearner Van der Laan et al. (2007) is an algorithm that constructs an ensemble of ML models predictions using a cross-validated metric and a set of normalized coefficients. By default Superlearner uses a ten-fold cross-validation procedure. It outputs a combination of user-provided ML models such that the RMSE of the final prediction is minimized by default Polley and Van der Laan (2010) (or any other user-defined metric defining the expected risk of the task at hand). In our setup, SuperLearner achieves this using 10fCV, where the training data is divided into ten equal portions or folds, the models are trained on nine folds, and the 10th fold is used as a test set. The models predict the target variable of the test set, and based on the RMSE of their predictions, SuperLearner assigns a coefficient. If an algorithm has a lower RMSE in 10fCV, it will be assigned a higher coefficient. Finally, it creates the ensemble as a linear combination of the constituent models multiplied by their respective coefficients. Note that this 10fCV is an internal procedure of model selection to build the SuperLearner ensemble model, and it is separate from the repeated cross-validation procedure which we described in **Section 3.1** and which is used to evaluate the performance and final results.

3.4 The Machine Learning Algorithms Used in Our Analysis

Following Dainotti et al. (2021) we analyze the coefficients assigned by SuperLearner to 12 ML algorithms, and pick those with a value greater than 0.05. In **Figure 2**, we show all the ML algorithms tested, and their coefficients. We pick the six algorithms above the 0.05 cutoff, which are: Enhanced

Adaptive Regression Through Hinge (EARTH), KSVM, Cforest, Ranger, Random Forest, and Linear Model. We provide brief explanations for each of them below.

Enhanced Adaptive Regression Through Hinges (EARTH) is an algorithm that allows for better modeling of predictor interaction and non-linearity in the data compared to the linear model. It is based on the Multivariate Adaptive Regression Splines method (MARS) Friedman and Roosen (1995). EARTH works by fitting a sum or product of hinges. Hinges are part-wise linear fits of the data that are joined such that the sum-of-squares residual error is minimized with each added term.

KSVM is an R implementation of the Support Vector Regression method (SVR). Similar to Support Vector Machine (SVM) Cortes and Vapnik (1995), SVR uses a kernel function to send its inputs to a higher-dimensional space where the data is linearly separable by a hyper-plane. SVR aims to fit this hyper-plane such that the prediction error is within a pre-specified threshold. For our purposes, KSVM uses the Gaussian kernel with the default parameters.

The Random Forest algorithm Breiman (2001); Ho (1995) seeks to extend decision trees capabilities by simultaneously generating multiple, independent decision trees. For regression tasks, Random Forest will return the average of the outputs of each of the generated decision trees. An advantage of Random Forest over decision trees is the reduction in the variance. However, Random Forest often suffers from low interpretability.

The Ranger algorithm is similar to Random Forest with the difference of extremely randomized trees (ERTs) Geurts et al. (2006) and quicker implementation.

Similar to Random Forest, the Cforest algorithm Hothorn et al. (2006) builds conditional inference trees that perform splits on significance tests instead of information gain.

We use the ordinary least squares (OLS) linear model found in the SuperLearner package. This model aims to minimize the mean squared error.

Note that we are using the default hyperparameter settings for all the algorithms.

3.5 Feature Engineering

Feature engineering is a broad term that incorporates two techniques: feature selection and feature creation. Feature selection is a method where the best predictors of a response variable are chosen from a larger pool of predictors. There exist multiple methods to perform feature selection. We are using the Least Absolute Selection and Shrinkage Operator (LASSO) method. Feature selection is an essential part of any ML study as it reduces the dimensionality of the data and minimizes the risk
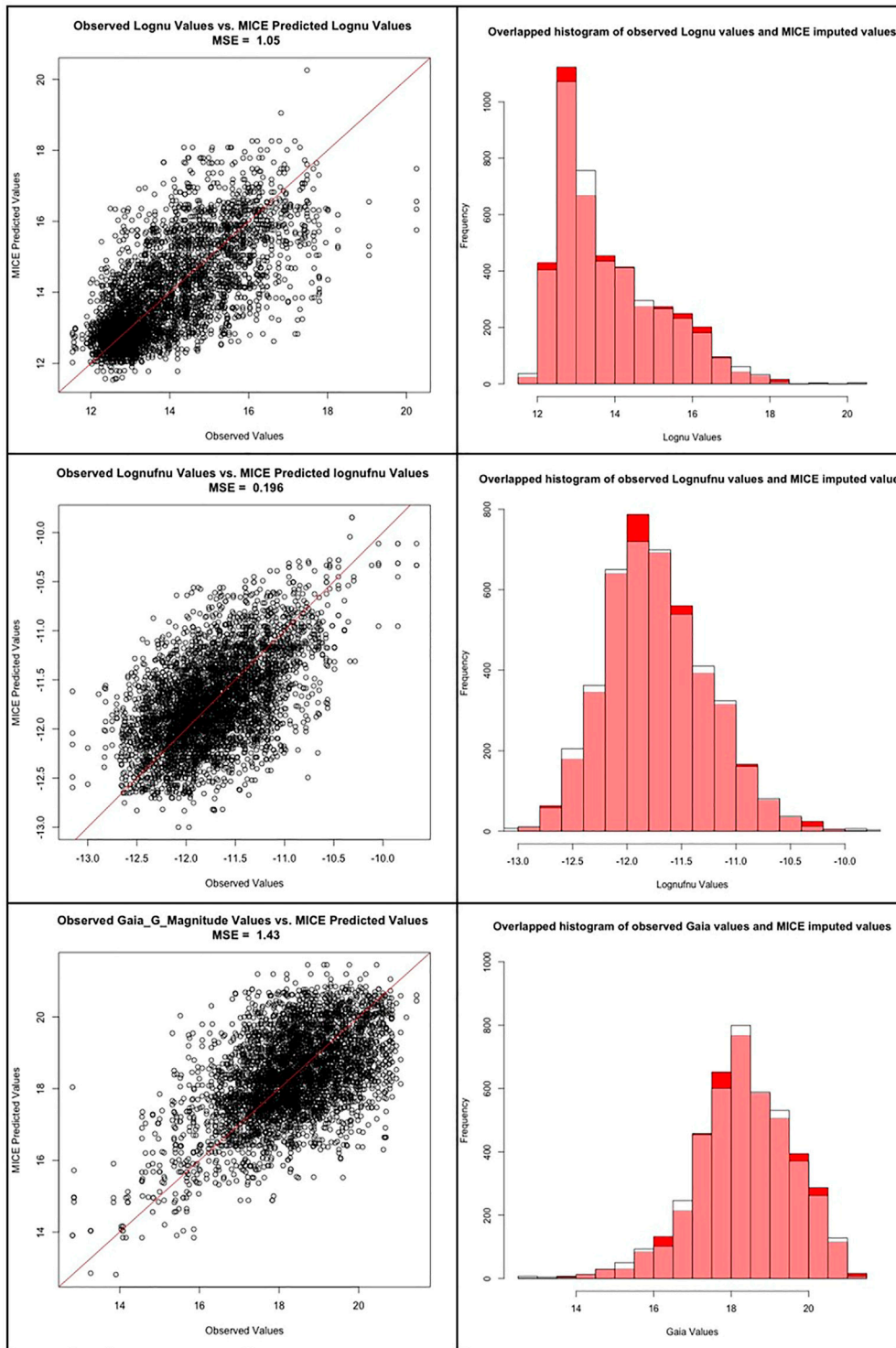
FIGURE 5 | First row: The scatter plot between the observed and predicted MICE values for the Logν predictor, followed by the overlapped histogram distributions of the same. Second row: The scatter plot between the observed and predicted MICE values for the Log$\nu f \nu$ predictor, followed by the overlapped histogram distributions of the same. Third row: The scatter plot between the observed and predicted MICE values for the *Gaia_G_Magnitude* predictor, followed by the overlapped histogram distributions of the same.
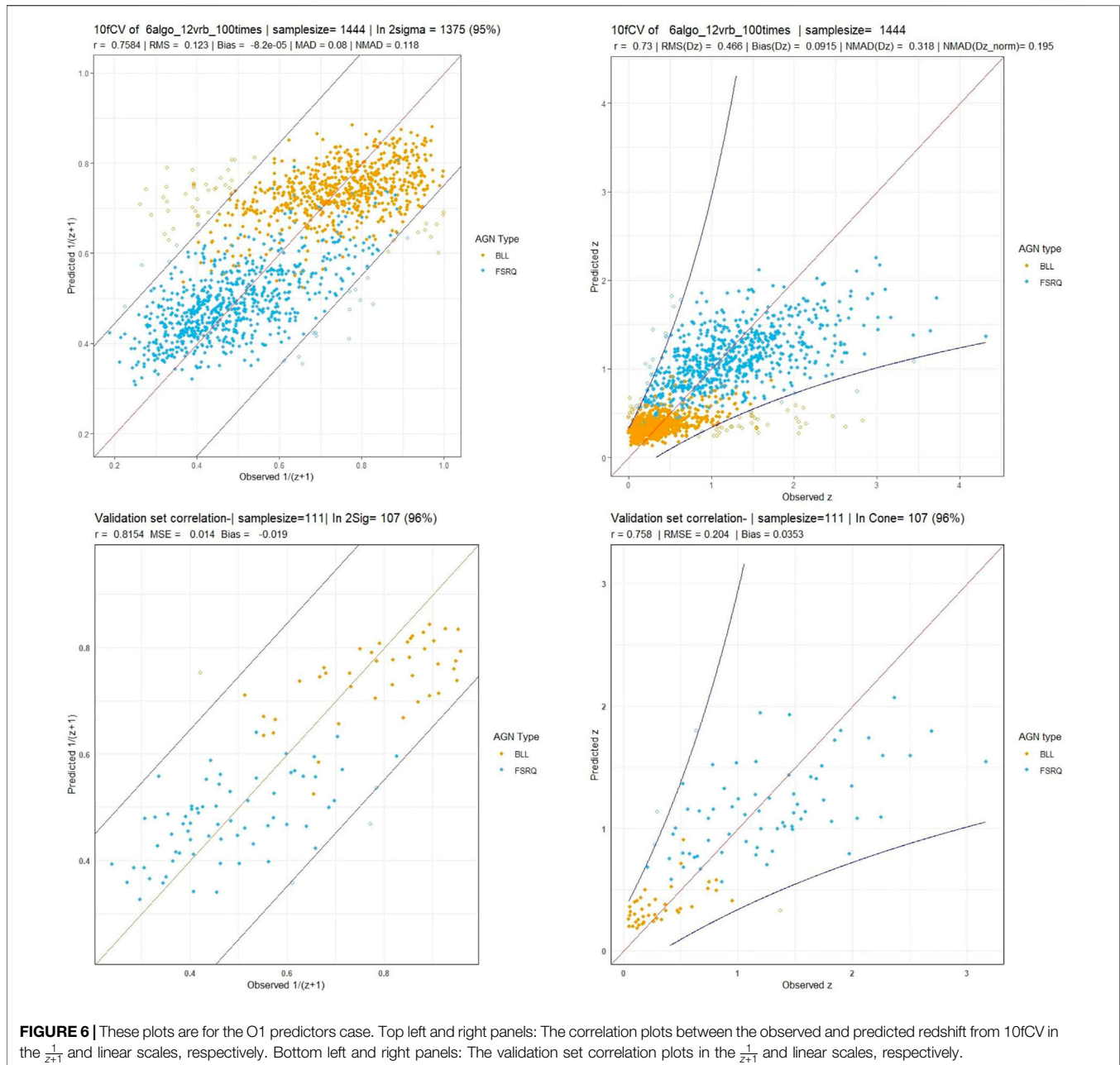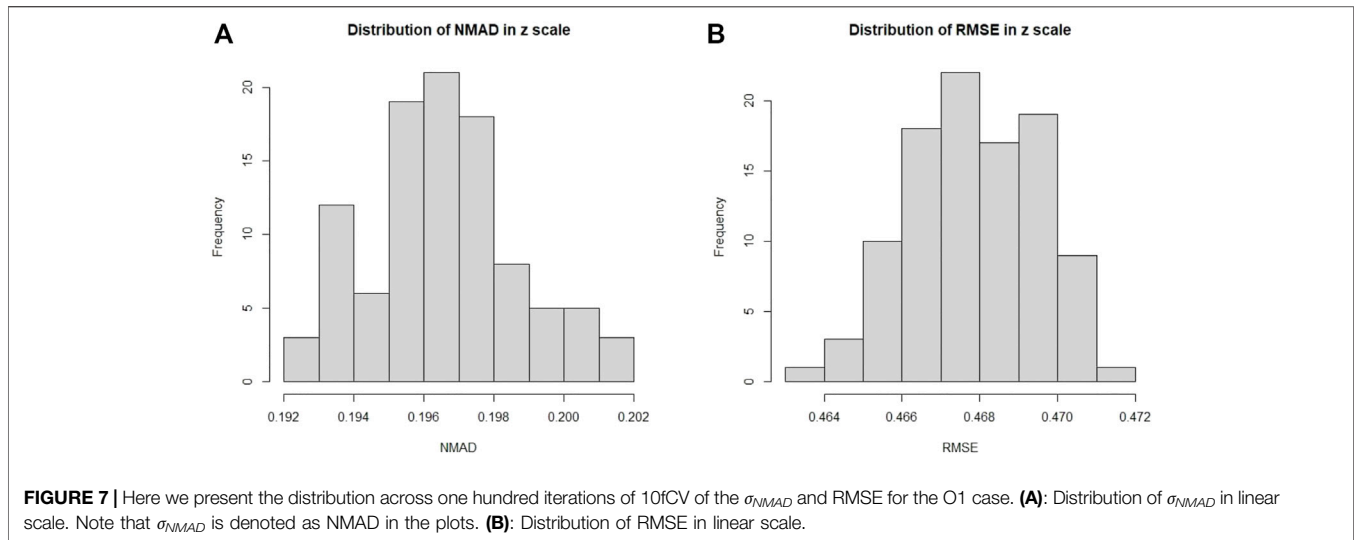
FIGURE 6 | These plots are for the O1 predictors case. Top left and right panels: The correlation plots between the observed and predicted redshift from 10fCV in the $\frac{1}{z+1}$ and linear scales, respectively. Bottom left and right panels: The validation set correlation plots in the $\frac{1}{z+1}$ and linear scales, respectively.

of overfitting. Feature creation is a technique where additional features are created from various combinations of existing properties. These combinations can be cross-multiplications, higher-order terms, or ratios. Feature creation can reveal hidden patterns in the data that ML algorithms might not be able to discern and consequently boost the performance.

In machine learning, some of the methods used by SuperLearner are linear by nature (BayesGLM, Lasso, elastic-net). Adding quadratic and multiplicative terms allows us to model some types of non-linear relationships. Interactions among variables are very important and can boost the prediction when used. The phrase "interaction among the variables" means the influence of one variable on the other;

however, not in an additive way, but rather in a multiplicative way. In our feature engineering procedure, we build these interactions by cross-products and squares of the initial variables. It is common that adding O2 predictors aids results since they may contain information not available in the O1 predictors.

In this study, we create 66 new features, which, as mentioned, are the cross-products and squares of the existing features of the 4LAC catalog. We denote the existing predictors of the 4LAC catalog as Order-1 (O1) predictors and the new predictors as Order-2 (O2). Thus, we expand the set of predictors from the initial eleven O1 predictors to a combined seventy-eight O1 and O2 predictors.

FIGURE 7 | Here we present the distribution across one hundred iterations of 10fCV of the σ_{NMAD} and RMSE for the O1 case. **(A)**: Distribution of σ_{NMAD} in linear scale. Note that σ_{NMAD} is denoted as NMAD in the plots. **(B)**: Distribution of RMSE in linear scale.

For features selection, LASSO Tibshirani (1996) is used. It works by constraining the ℓ^1 norm of the coefficient vector to be less than or equal to a tuning parameter λ while fitting a linear model to the data. The predictors that LASSO chooses have a non-zero coefficient for the largest λ value with the property that the corresponding prediction error is within one standard deviation of the minimum prediction error Friedman et al. (2010); Birnbaum (1962); Hastie and Tibshirani. (1987), Hastie and Tibshirani. (1990); Friedman et al. (2010). This study performs LASSO feature selection on a fold-by-fold basis during external 10fCV. Optimal features are picked using LASSO for nine of the ten folds, and the predictions on the 10th fold are performed using these selected features. This step is iterated such that for every combination of nine folds, an independent set of features is picked. This usage of LASSO is in contrast to Dainotti et al. (2021), where the best features are picked for the entire training set. Our updated technique ensures that during the 10fCV, LASSO only picks the best predictors based on the training data, and the test set does not affect the models. This feature selection method is applied to both the O1 and O2 predictor sets.
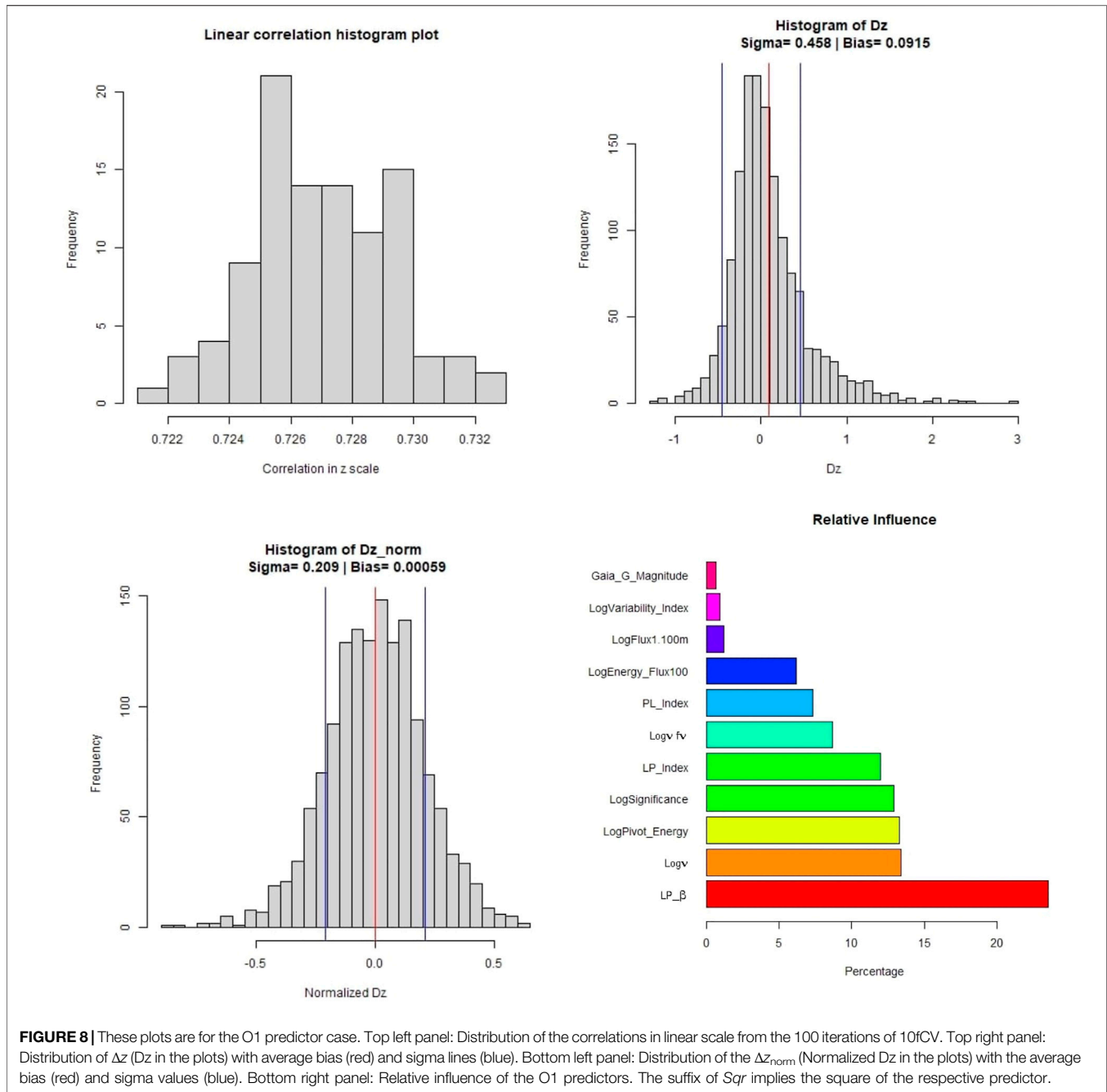
4 RESULTS

The quality of the MICE imputations depends on the information density of the entire data set. Hence, to ensure the best possible imputations we use all 1897 AGNs which remain after the removal of outliers and non-BLL and non-FSRQ AGNs. The pattern of the missing entries in our data set is shown in **Figure 3**, and they are present in only three predictors, namely, Logν, Log$\nu f\nu$, and *Gaia_G_Magnitude* (see **Sec. 2**). There are 237 AGNs which have missing values in both Logν and Log$\nu f\nu$, and 343 AGNs have a missing value in *Gaia_G_Magnitude*. MICE is used to fill the missing values of these AGNs. In **Figure 4** we show the distributions of Logν, Log$\nu f\nu$, and *Gaia_G_Magnitude* with and without MICE. The quality of the MICE imputations can be

evaluated in part by comparing the original distribution of a variable and its distribution with imputations. If the imputations alter the distribution, the results cannot be trusted and would require additional precautions or measures to deal with the missing values. However, as can be discerned from the plots (**Figure 4**), the MICE imputations are indeed following the underlying distribution for the three predictors, and hence we confidently incorporate them into our analysis. We impute 465 data points, 24% of our data set, resulting in a training sample of 1,444 AGNs and a generalization sample of 453 AGNs. The two sets are detailed in **Table. 1**.

4.1 Multivariate Imputation by Chained Equations Reliability Analysis

In the work by Luken et al. (2021), they present an extensive analysis of the reliability of MICE imputations. However, since they use a different dataset than ours, a similar investigation regarding the performance of MICE is essential. Thus, we take 1,432 AGNs from our catalog with no missing entries and randomly dropped 20% of the entries from each of the three predictors which have missing entries, namely: Logν, Log$\nu f\nu$, and *Gaia_G_Magnitude*. We then impute these dropped entries using MICE, as described in **Section 3.2**. This process is repeated fifteen times, and each time a different set of random entries are dropped. Furthermore, as we can see in **Figure 5**, the observed vs predicted values for Logν, Log$\nu f\nu$ and *Gaia_G_Magnitude* are concentrated about the $y = x$ line, with little variance. The mean squared error (MSE), defined as the, of the observed values vs the MICE imputed values for Logν, Log$\nu f\nu$, and *Gaia_G_Magnitude* were 1.05, 0.196, and 1.43, respectively. Thus, the MSEs are all small, which provides evidence that the MICE imputed effectively. Note that if MICE imputes effectively, then the imputed values and observed values should come from the same distribution for each of the three variables. To check this, we performed a Kolmogorov-Smirnov (KS) test on the observed vs MICE imputed values for each of the three
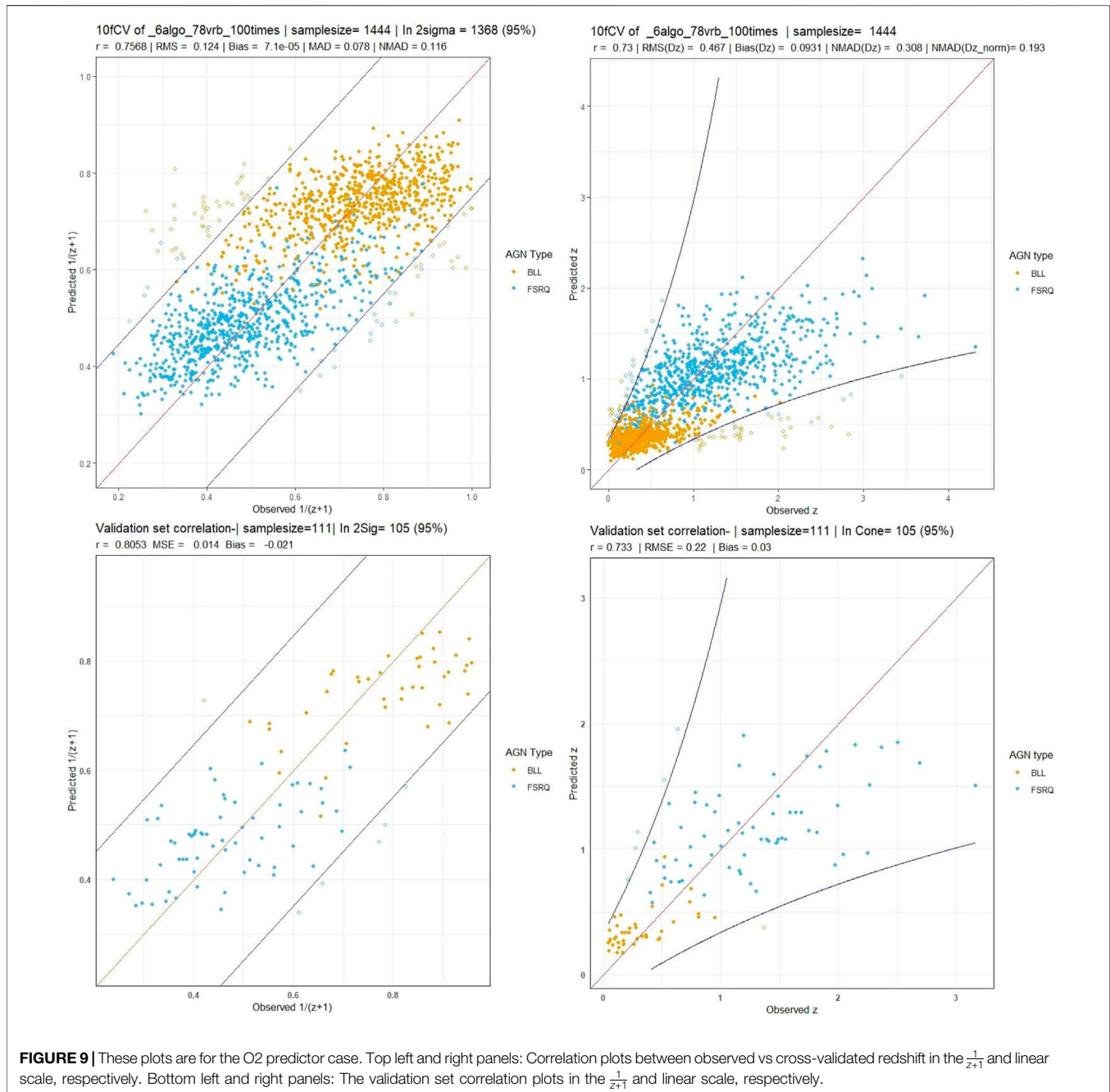
FIGURE 8 | These plots are for the O1 predictor case. Top left panel: Distribution of the correlations in linear scale from the 100 iterations of 10fCV. Top right panel: Distribution of Δz (Dz in the plots) with average bias (red) and sigma lines (blue). Bottom left panel: Distribution of the Δz_{norm} (Normalized Dz in the plots) with the average bias (red) and sigma values (blue). Bottom right panel: Relative influence of the O1 predictors. The suffix of *Sqr* implies the square of the respective predictor.

variables with missing entries. The *p*-values of the KS test for Log*v*, Log*vfv*, and *Gaia_G_Magnitude* were 0.744, 0.5815, and 0.6539, respectively. Since each of these *p*-values is above 0.05, we cannot reject the null-hypothesis; namely, we conclude that the observed values and the MICE imputed values come from same distributions for any of the three variables. As shown in **Figure 5**, the overlapped histogram of the observed vs MICE imputed values for Log*v*, Log*vfv* and *Gaia_G_Magnitude* are each very similar, which reinforces the findings of the KS test - namely, that they are from the same distribution. This provides additional proof for the accuracy, and reliability of the MICE imputations.

4.2 With O1 Variables

The O1 variable set consists of 12 predictors, including the categorical variable *LabelNo*, which distinguishes between BLLs and FSRQs. LASSO chooses the best predictors from within this set for each fold in the 10fCV as explained in **Section 3.5**.
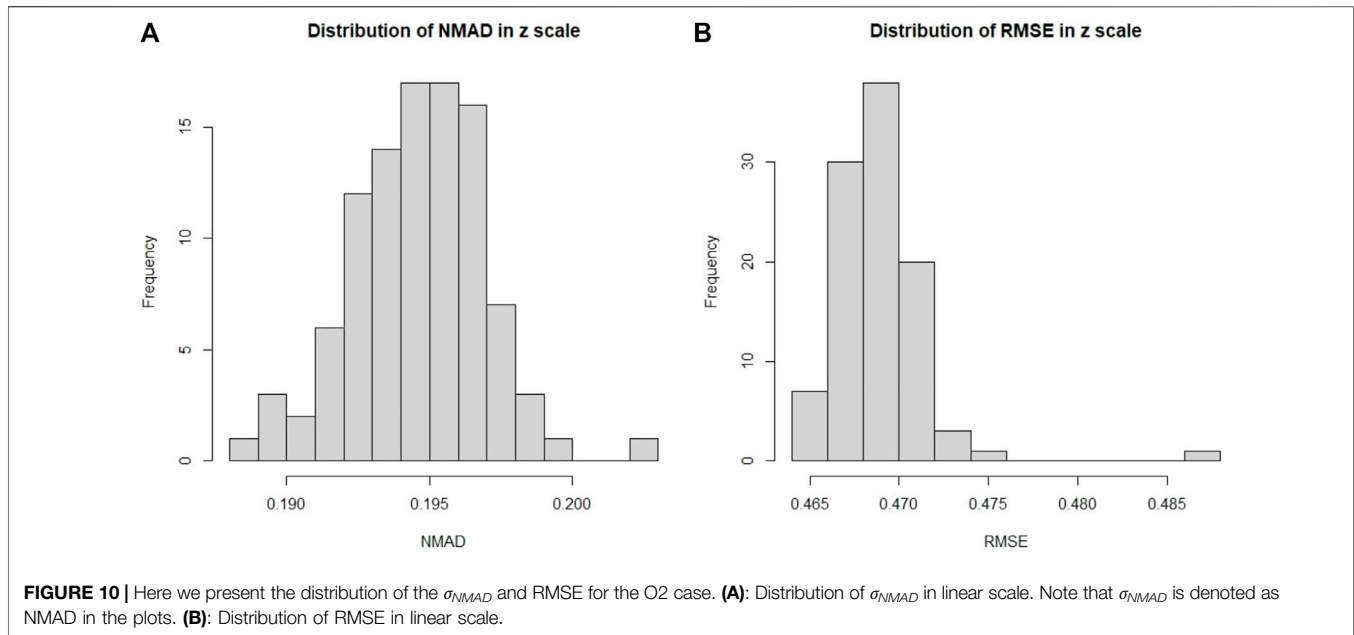
Using this feature set with the six algorithms mentioned, we obtain a correlation in the $1/(z + 1)$ scale of 75.8%, a σ of 0.123, an RMSE of 0.123, and a σ_{NMAD} of 0.118. In the linear redshift scale (z scale), we obtain a correlation of 73%, an RMSE of 0.466, a σ of 0.458, a bias of 0.092, and a σ_{NMAD} of 0.318. In the normalized

FIGURE 9 | These plots are for the O2 predictor case. Top left and right panels: Correlation plots between observed vs cross-validated redshift in the $\frac{1}{z+1}$ and linear scale, respectively. Bottom left and right panels: The validation set correlation plots in the $\frac{1}{z+1}$ and linear scale, respectively.

scale (Δz_{norm}), the RMSE obtained is 0.209, bias is 6×10^{-3}, and σ_{NMAD} is equal to 0.195. The correlation plots are shown in **Figure 6**, with the left panel showing the correlation in the $1/(z + 1)$ scale and the right panel showing the correlation in the z scale. We obtain a low 5% catastrophic outlier percentage in this scenario. The lines in blue depict the 2σ curves for each plot, where the σ is calculated in the $1/(z + 1)$ scale.

In **Figure 7**, we present the distributions of σ_{NMAD} and RMSE across the one hundred iterations. Note that σ_{NMAD} is written as NMAD in the plots for brevity.

In **Figure 8**, we present the distributions of various parameters and the normalized relative influence plot of the 11 predictors - *LabelNo* is excluded, as its a categorical variable. The top left panel shows the variation in the linear correlation obtained from the one hundred iterations. The top right panel shows the distribution of Δz along with the σ (blue vertical line) and bias (red vertical line) values. The bottom left panel shows the distribution of the Δz_{norm} along with the bias and σ presented similarly. Finally, the barplot in the bottom right panel shows the relative influence of the 11 predictors used. *LP_β* has the highest

FIGURE 10 | Here we present the distribution of the σ_{NMAD} and RMSE for the O2 case. **(A)**: Distribution of σ_{NMAD} in linear scale. Note that σ_{NMAD} is denoted as NMAD in the plots. **(B)**: Distribution of RMSE in linear scale.

influence, followed by *Logν*, *LogPivot_Energy*, and *LogSignificance*. Surprisingly, *Gaia_G_Magnitude* has the least influence at $\approx 1\%$, in contrast to Dainotti et al. (2021), where we found it to be quite significant at $\approx 11\%$ influence. The difference we obtain from this analysis and the previous one of Dainotti et al. (2021) lies in the data set and that MICE had not been used.

4.3 With O2 Variables

The O2 variables, 78 in total, are made from cross-products of the O1 variables. As in the O1 case, LASSO feature selection is performed on a fold-by-fold basis, after which the SuperLearner ensemble with the six algorithms previously mentioned is trained and makes predictions. The cross-validation and validation correlation plots are presented in **Figure 9**.

As shown in the previous section, we have correlation plots in the $1/(z + 1)$ scale and the z scale. In the $1/(z + 1)$ scale, we get a correlation of 75.6%, RMSE of 0.124, and σ_{NMAD} of 0.116. In the z scale, we obtain a correlation of 73%, RMSE of 0.467, and σ_{NMAD} of 0.308. We obtain the statistical parameters for Δz: an RMSE of 0.467, a σ of 0.458, a bias of 0.093, and a σ_{NMAD} of 0.308. For Δz_{norm}, we obtain an RMSE of 0.21, a bias of 7×10^{-4}, and a σ_{NMAD} of 0.193. We have a similar catastrophic outlier percentage (5%) as the O1 variable case, although the number of AGNs predicted outside the 2σ cone is seven AGNs more. This discrepancy can be attributed to the randomness inherent in our calculations and additional noise introduced by the O2 predictors.

In **Figure 10**, we show the distributions of σ_{NMAD} and RMSE. Note that there is an outlier during the analysis, which leads to the unusually high RMSE value seen in the distribution.

Figure 11 shows the distribution plots for various parameters. The top left panel shows the distribution of the correlations across the one hundred iterations. There is an outlier in the distribution
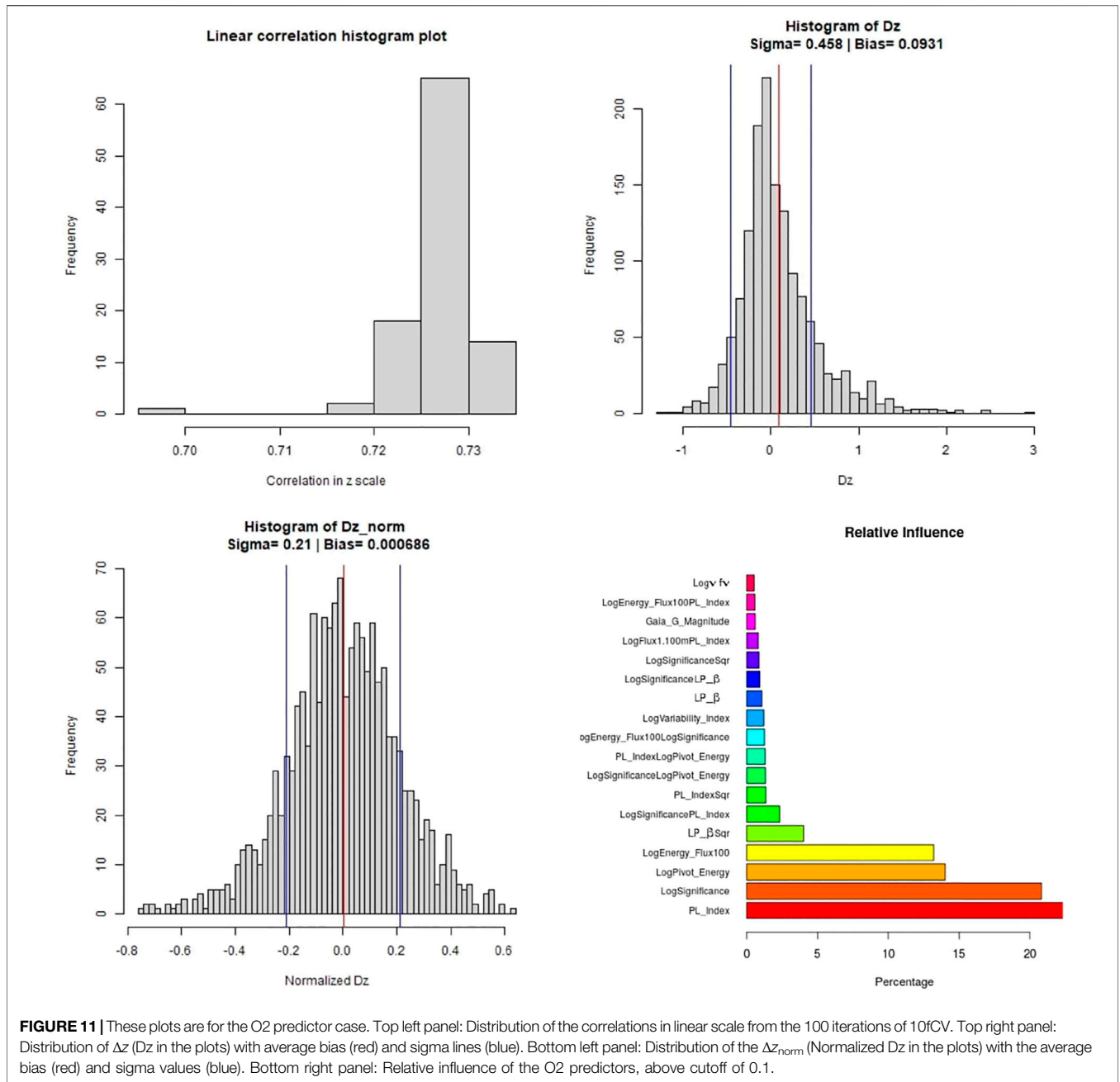
of the correlation plot, corresponding to the distributions of RMSE in **Figure 10**. This scenario only happens with the O2 variable set and with MICE imputations. Apart from this fluctuation, most of the correlations lie around 73%. The histogram distribution plots for Δz (top right) and Δz_{norm} (bottom left) show a similar spread as in the case of the O1 variable set. We only present predictors with influence greater than 0.5% in the relative influence plot. In this case, *PL_Index* turns out to have the highest influence, over 20%, followed by *LogSignificance*, *LogPivot_Energy*, and *LogEnergy_Flux*100.

We note that out of the 11 O1 predictors with relative influences, only 3 have less than 5% influence, and out of the 78 O2 predictors, only 4 have greater than 5% influence. Thus, the majority of the O2 predictors do not seem to provide much additional information about the redshift.

In **Tables 2** and **3** we provide a comparision between the results obtained in the two experiments we have here with MICE, and one without MICE imputations. The latter results have been taken from Narendra et al. (2022).

5 DISCUSSIONS AND CONCLUSION

In Dainotti et al. (2021), the correlation between the observed and predicted redshift achieved with a training set of 730 AGNs was 71%, with RMSE of 0.434, σ_{NMAD} (Δz_{norm}) of 0.192, and a catastrophic outlier of 5%. Here, with the use of an updated 4LAC catalog, O1 predictors, and the MICE imputation technique, along with additional ML algorithms in the SuperLearner ensemble, we achieve a correlation of 73% between the observed and predicted redshift, an RMSE of 0.466, σ_{NMAD} (Δz_{norm}) of 0.195 and a catastrophic outlier of 5%. Although the RMSE and σ_{NMAD} (Δz_{norm}) are increasing by 7 and 1.5%, respectively, we are able to maintain the

FIGURE 11 | These plots are for the O2 predictor case. Top left panel: Distribution of the correlations in linear scale from the 100 iterations of 10fCV. Top right panel: Distribution of Δz (Dz in the plots) with average bias (red) and sigma lines (blue). Bottom left panel: Distribution of the Δz_{norm} (Normalized Dz in the plots) with the average bias (red) and sigma values (blue). Bottom right panel: Relative influence of the O2 predictors, above cutoff of 0.1.

catastrophic outliers at 5%, while increasing the correlation by 3%. These results are achieved with a data sample 98% larger than the one used by Dainotti et al. (2021). Note that this achievement is not trivial, as a larger data set does not guarantee favourable results.

With the O2 predictor set, we obtain a similar correlation of 73% between the predicted and observed redshifts. However, compared to the O1 case, the RMSE goes up by 0.2%–0.467 and the σ_{NMAD} (Δz_{norm}) goes down by 1% to 0.193. The catastrophic outlier percentage is maintained at 5% in both cases.

The most influential O1 predictors in this study were LP_β, $Log\nu$, $LogPivot\_Energy$, $LogSignificance$, LP_Index, PL_Index,

and $LogEnergy\_Flux$, each of which has a relative influence greater than 5%. LP_β was also the most influential predictor in Dainotti et al. (2021), followed by $LogPivot\_Energy$, $LogSignificance$, $LogEnergy\_Flux$, and $Log\nu$. The main difference in the relative influences of the predictors in these studies is that in the O1 case with MICE, LP_Index and PL_Index are the 5th and 7th most influential predictors, respectively, while in Dainotti et al. (2021), they were not influential.

Among the O2 predictors, PL_Index is the most influential, followed by $LogSignificance$, $LogPivot\_Energy$, and $LogEnergy\_Flux$, each of which has a relative influence

TABLE 2 | Comparison of the statistical metrics across the different experiments performed. These have been calculated for the $1/(z + 1)$ scale.

Metric	SL with O1	SL with O2	Without MICE
r	0.758	0.757	0.781
RMSE	0.123	0.124	0.119
Bias	-8.2×10^{-5}	7.1×10^{-5}	4×10^{-4}
σ_{NMAD}	0.118	0.116	0.113
σ	0.209	0.210	0.119

TABLE 3 | Comparision of the statistical metrics across the different experiments performed. These have been calculated for the z scale.

Metric	SL with O1	SL with O2	Without MICE
r	0.73	0.73	0.74
RMSE (Δz)	0.466	0.467	0.467
Bias (Δz)	0.0915	0.0931	0.095
σ_{NMAD} (Δz)	0.318	0.308	0.321
σ (Δz)	0.458	0.458	0.458
Bias (Δz_{norm})	$5.9, \times, 10^{-4}$	$6.9, \times, 10^{-4}$	9.6×10^{-4}
σ_{NMAD} (Δz_{norm})	0.195	0.193	0.195
σ (Δz_{norm})	0.209	0.210	0.208

greater than 5%. Note that the only O2 predictors with influence greater than 5% are those we have just listed and they are also O1 predictors. When additional variables are added it is not guaranteed that the most influential variables will be kept the same. This is true for both parametric and non-parametric models. The influence is a measure of how much your improvement in the prediction changes when you remove one variable in relation to the presence of the other variables. Thus, these measures depend on the other variables in the model and are different when O2 variables are added. We can conclude from these results that the O1 predictors contain most of the predictive information for redshift, in the case of the 4LAC catalog. Furthermore, we note that obtaining results with the O2 set takes more time than with the O1 set due to the larger list of predictors. However, in other catalogs, such O2 predictors might perform better and be an avenue worth exploring in the future.

Here, we use MICE on the O1 variables, because this allows MICE to act on three variables which present missing entries. In this way, we can control the effectiveness of MICE and the results. We agree with the referee that imputing the MICE in the cross products would imply an imputation on variables that are currently not defined and most importantly would allow more uncertainty when the cross products would involve for example two variables with missing entries. If we had used MICE in the O2 parameters we would have had a large number of imputation which would be less controllable. From these results, we can discern that the MICE imputation technique is a robust method to mitigate the issue of missing entries in a catalog while maintaining the predictive power of the data.

AUTHOR CONTRIBUTIONS

SG and AN were responsible for improving parts of the initial code, the generation of the final results, and the writing of the manuscript. This project was initiated by MD. The original code was developed by her and MB. It was later modified to suit the specific needs of this project. MD also participated in all the discussions regarding the project was instrumental in shaping the outcome, and determining the structure of the manuscript. MB helped with the statistical results of the work. Dr. Agniezska Pollo participated in crucial discussions regarding the direction of the research. AP helped in editing the manuscript along with discussions about the results. ER helped in editing the structure of the manuscript along with discussions about the procedures that were adopted along with the idea of using LASSO feature selection inside the cross-validation. IL helped us in acquiring and explaining the data and its features, along with corresponding labels of BLL and FSRQ. He also helped in making the manuscript clearer.

REFERENCES

Abdollahi, S., Acero, F., Ackermann, M., Ajello, M., Atwood, W. B., Axelsson, M., et al. (2020). Fermi Large Area Telescope Fourth Source Catalog. *ApJS* 247, 33. doi:10.3847/1538-4365/ab6bcb

Ackermann, M., Ajello, M., Albert, A., Atwood, W. B., Baldini, L., Ballet, J., et al. (2015). Multiwavelength Evidence for Quasi-Periodic Modulation in the Gamma-Ray Blazar PG 1553+113. *Astrophysical J. Lett.* 813, L41. doi:10.1088/2041-8205/813/2/L41

Ackermann, M., Ajello, M., Allafort, A., Baldini, L., Ballet, J., Bastieri, D., et al. (2012). GeV Observations of Star-forming Galaxies with the Fermi Large Area Telescope. *Astrophysical J.* 755, 164. doi:10.1088/0004-637X/755/2/164

Aihara, H., Prieto, C. A., An, D., Anderson, S. F., Aubourg, É., Balbinot, E., et al. (2011). Erratum: The Eighth Data Release of the sloan Digital Sky Survey: First Data from SDSS-III. *Astrophysical J. Suppl. Ser.* 193, 29. doi:10.1088/0067-0049/195/2/26

Ajello, M., Angioni, R., Axelsson, M., Ballet, J., Barbiellini, G., Bastieri, D., et al. (2020). The Fourth Catalog of Active Galactic Nuclei Detected by the Fermi Large Area Telescope. *ApJ* 892, 105. doi:10.3847/1538-4357/ab791e

Birnbaum, A. (1962). On the Foundations of Statistical Inference. *J. Am. Stat. Assoc.* 57, 269–306. doi:10.1080/01621459.1962.10480660

Breiman, L. (2001). Random Forests. *Machine Learn.* 45, 5–32. doi:10.1023/a:1010933404324

Brescia, M., Cavuoti, S., D'Abrusco, R., Longo, G., and Mercurio, A. (2013). Photometric Redshifts for Quasars in Multi-Band Surveys. *ApJ* 772, 140. doi:10.1088/0004-637x/772/2/140

Brescia, M., Salvato, M., Cavuoti, S., Ananna, T. T., Riccio, G., LaMassa, S. M., et al. (2019). Photometric Redshifts for X-ray-selected Active Galactic Nuclei in the eROSITA Era. *Monthly Notices R. Astronomical Soc.* 489, 663–680. doi:10.1093/mnras/stz2159

Cavuoti, S., Brescia, M., D'Abrusco, R., Longo, G., and Paolillo, M. (2014). Photometric Classification of Emission Line Galaxies with Machine-Learning Methods. *Monthly Notices R. Astronomical Soc.* 437, 968–975. doi:10.1093/mnras/stt1961

Chiang, J., Fichtel, C. E., Von Montigny, C., Nolan, P. L., and Petrosian, V. (1995). The Evolution of Gamma-Ray–loud Active Galactic Nuclei. *ApJ* 452, 156. doi:10.1086/176287

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Machine Learn.* 20, 273–297. doi:10.1023/a:1022627411411

Curran, S. J. (2020). QSO Photometric Redshifts from SDSS, WISE, and GALEX Colours. *Monthly Notices R. Astronomical Soc. Lett.* 493, L70–L75. doi:10.1093/mnrasl/slaa012

Dainotti, M. G., Bogdan, M., Narendra, A., Gibson, S. J., Miasojedow, B., Liodakis, I., et al. (2021). Predicting the Redshift of γ-Ray-loud AGNs Using Supervised Machine Learning. *ApJ* 920, 118. doi:10.3847/1538-4357/ac1748

D'Isanto, A., and Polsterer, K. L. (2018). Photometric Redshift Estimation via Deep Learning. Generalized and Pre-Classification-Less, Image Based, Fully Probabilistic Redshifts. *aap* 609, A111. doi:10.1051/0004-6361/201731326

Domínguez, A., Wojtak, R., Finke, J., Ajello, M., Helgason, K., Prada, F., et al. (2019). A New Measurement of the Hubble Constant and Matter Content of the Universe Using Extragalactic Background Light γ-ray Attenuation. *ApJ* 885, 137. doi:10.3847/1538-4357/ab4a0e

Fermi-LAT CollaborationAbdollahi, S., Ackermann, M., et al. (2018). *Science* 362, 1031.

Fotopoulou, S., and Paltani, S. (2018). CPz: Classification-Aided Photometric-Redshift Estimation. *A&A* 619, A14. doi:10.1051/0004-6361/201730763

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* 33, 1. doi:10.18637/jss.v033.i01

Friedman, J. H., and Roosen, C. B. (1995). An Introduction to Multivariate Adaptive Regression Splines. *Stat. Methods Med. Res.* 4, 197–217. doi:10.1177/096228029500400303

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely Randomized Trees. *Machine Learn.* 63, 42–63. doi:10.1007/s10994-006-6226-1

Hastie, T. J., and Tibshirani, R. J. (1990). *Generalized Additive Models*, 43. Boca Raton: CRC Press.

Hastie, T., and Tibshirani, R. (1987). Generalized Additive Models: Some Applications. *J. Am. Stat. Assoc.* 82, 371–386. doi:10.1080/01621459.1987.10478440

Hildebrandt, H., Arnouts, S., Capak, P., Moustakas, L. A., Wolf, C., Abdalla, F. B., et al. (2010). PHAT: PHoto-zAccuracy Testing. *A&A* 523, A31. doi:10.1051/0004-6361/201014885

Ho, T. K. (1995). "Random Decision Forests," in Proceedings of the Third International Conference on Document Analysis and Recognition Volume 1, ICDAR '95, Montreal, QC, Canada, August 14-16, 1995 (USA: IEEE Computer Society), 278.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *J. Comput. Graphical Stat.* 15, 651–674. doi:10.1198/106186006x133933

Ilbert, O., Capak, P., Salvato, M., Aussel, H., McCracken, H., Sanders, D. B., et al. (2008). Cosmos Photometric Redshifts with 30-Bands for 2-deg^2. *Astrophysical J.* 690, 1236–1249. doi:10.1088/0004-637X/690/2/1236

Jones, E., and Singal, J. (2017). Analysis of a Custom Support Vector Machine for Photometric Redshift Estimation and the Inclusion of Galaxy Shape Information. *A&A* 600, A113. doi:10.1051/0004-6361/201629558

Liodakis, I., Hovatta, T., Huppenkothen, D., Kiehlmann, S., Max-Moerbeck, W., and Readhead, A. C. S. (2018). Constraining the Limiting Brightness Temperature and Doppler Factors for the Largest Sample of Radio-Bright Blazars. *ApJ* 866, 137. doi:10.3847/1538-4357/aae2b7

Liodakis, I., Pavlidou, V., Hovatta, T., Max-Moerbeck, W., Pearson, T. J., Richards, J. L., et al. (2017). Bimodal Radio Variability in OVRO-40 M-Monitored Blazars. *MNRAS* 467, 4565–4576. doi:10.1093/mnras/stx432

Little, R. J., and Rubin, D. B. (2019). *Statistical Analysis with Missing Data*, 793. Hoboken: John Wiley & Sons.

Logan, C. H. A., and Fotopoulou, S. (2020). Unsupervised star, Galaxy, QSO Classification. *A&A* 633, A154. doi:10.1051/0004-6361/201936648

Luken, K. J., Padhy, R., and Wang, X. R. (2021). Missing Data Imputation for Galaxy Redshift Estimation. *arXiv:2111.13806.*

Marcotulli, L., Ajello, M., and Di Mauro, M. (2020). "The Density of Blazars above 100 MeV and the Origin of the Extragalactic Gamma-ray Background," in American Astronomical Society Meeting Abstracts (American Astronomical Society Meeting Abstracts), 405–406.235

Nakoneczny, S. J., Bilicki, M., Solarz, A., Pollo, A., Maddox, N., Spiniello, C., et al. (2019). Catalog of Quasars From the Kilo-Degree Survey Data Release 3. *aap* 624, A13. doi:10.1051/0004-6361/201834794

Narendra, A., Gibson, S. J., Dainotti, M. G., Bogdan, M., Pollo, A., Liodakis, I., et al. (2022). Predicting the Redshift of Gamma-ray Loud AGNs Using Supervised Machine Learning: Part 2. *arXiv:2201.05374.*

Pasquet-Itam, J., and Pasquet, J. (2018). Deep Learning Approach for Classifying, Detecting and Predicting Photometric Redshifts of Quasars in the Sloan Digital Sky Survey Stripe 82. *A&AAstronomy & Astrophysics* 611, A97. doi:10.1051/0004-6361/201731106

Petrosian, V. (1976). Surface Brightness and Evolution of Galaxies. *ApJ* 209, L1. doi:10.1086/182253

Polley, E. C., and Van der Laan, M. J. (2010). Super Learner in Prediction. U.C. Berkeley Division of Biostatistics Working Paper Series. Bepress. Available at: https://biostats.bepress.com/ucbbiostat/paper266

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika* 63, 581–592. doi:10.1093/biomet/63.3.581

Salvato, M., Ilbert, O., and Hoyle, B. (2019). The many Flavours of Photometric Redshifts. *Nat. Astron.* 3, 212–222. doi:10.1038/s41550-018-0478-0

Schafer, J. L., and Graham, J. W. (2002). Missing Data: Our View of the State of the Art. *Psychol. Methods* 7, 147–177. doi:10.1037/1082-989x.7.2.147

Singal, J. (2015). A Determination of the Gamma-ray Flux and Photon Spectral index Distributions of Blazars from theFermi-LAT 3LAC. *Mon. Not. R. Astron. Soc.* 454, 115–122. doi:10.1093/mnras/stv1964

Singal, J., Ko, A., and Petrosian, V. (2014). Gamma-Ray Luminosity and Photon Index Evolution of FSRQ Blazars and Contribution to the Gamma-Ray Background. *Astrophysical J.* 786, 109.

Singal, J., Ko, A., and Petrosian, V. (2013a). Flat Spectrum Radio Quasar Evolution and the Gamma-ray Background. *Proc. IAU* 9, 149–152. doi:10.1017/s1743921314003597

Singal, J., Petrosian, V., and Ajello, M. (2012). Flux and Photon Spectral Index Distributions Offermi-Lat Blazars and Contribution to the Extragalactic Gamma-ray Background. *ApJ* 753, 45. doi:10.1088/0004-637x/753/1/45

Singal, J., Petrosian, V., and Ko, A. (2013b). Cosmological Evolution of the FSRQ Gamma-ray Luminosity Function and Spectra and the Contribution to the Background Based on Fermi-LAT Observations. *AAS/High Energy Astrophysics Division#* 13, 300–307.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B (Methodological)* 58, 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x

Van Buuren, S., and Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* 45, 1–67. doi:10.18637/jss.v045.i03

Van der Laan, M. J., Polley, E. C., and Hubbard, A. E. 2007, Super learner. *Stat. Appl. Genet. Mol. Biol.*, 6. doi:10.2202/1544-6115.1309

Venters, T. M., and Pavlidou, V. (2013). Probing the Intergalactic Magnetic Field with the Anisotropy of the Extragalactic Gamma-ray Background. *MNRAS* 432, 3485–3494. doi:10.1093/mnras/stt697

Wakely, S. P., and Horan, D. (2008). "TeVCat: An online catalog for Very High Energy Gamma-Ray Astronomy," in International Cosmic Ray Conference, 1341–1344.3

Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., Ressler, M. E., Cutri, R. M., Jarrett, T., et al. (2010). The Wide-field Infrared Survey Explorer (WISE): Mission Description and Initial On-orbit Performance. *Astronomical J.* 140, 1868–1881. doi:10.1088/0004-6256/140/6/1868

Yang, Q., Wu, X.-B., Fan, X., Jiang, L., McGreer, I., Green, R., et al. (2017). Quasar Photometric Redshifts and Candidate Selection: A New Algorithm Based on Optical and Mid-infrared Photometric Data. *Astronomical J.* 154, 269. doi:10.3847/1538-3881/aa943c

Zhang, K., Schlegel, D. J., Andrews, B. H., Comparat, J., Schäfer, C., Vazquez Mata, J. A., et al. (2019). Machine-learning Classifiers for Intermediate Redshift Emission-Line Galaxies. *ApJ* 883, 63. doi:10.3847/1538-4357/ab397e