Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **525,** 1731–1745 (2023) https://doi.org/10.1093/mnras/stad2193
Advance Access publication 2023 August 18

Fermi LAT AGN classification using supervised machine learning

Nathaniel Cooper [1]★ Maria Giovanna Dainotti,[2,3,4] Aditya Narendra,[5,6] Ioannis Liodakis [7] and Malgorzata Bogdan[8,9]

[1]*United States Merchant Marine Academy, Kings Point, NY 11024, USA*
[2]*National Astronomical Observatory of Japan, Mitaka 181-8588, Japan*
[3]*Space Science Institute, 4750 Walnut St, Suite 205, Boulder, CO 80301, USA*
[4]*School of Physical Sciences, The Graduate University for Advanced Studies, Shonankokusaimura, Hayama, Miura District, Kanagawa 240-0193, Japan*
[5]*Astronomical Observatory of Jagiellonian University, PL-30-244 Kraków, Poland*
[6]*Doctoral School of Exact and Natural Sciences, Jagiellonian University, PL-31-007 Kraków, Poland*
[7]*Finnish Centre for Astronomy with ESO (FINCA), University of Turku, FI-20014 Turku, Finland*
[8]*Department of Mathematics, University of Wroclaw, PL-50-370 Wroclaw, Poland*
[9]*Department of Statistics, Lund University, SE-220 07 Lund, Sweden*

Accepted 2023 July 16. Received 2023 June 26; in original form 2022 December 22

ABSTRACT

Classifying active galactic nuclei (AGNs) is a challenge, especially for BL Lacertae objects (BLLs), which are identified by their weak emission line spectra. To address the problem of classification, we use data from the fourth *Fermi* Catalog, Data Release 3. Missing data hinder the use of machine learning to classify AGNs. A previous paper found that Multivariate Imputation by Chain Equations (MICE) imputation is useful for estimating missing values. Since many AGNs have missing redshift and the highest energy, we use data imputation with MICE and *k*-nearest neighbours (kNN) algorithm to fill in these missing variables. Then, we classify AGNs into the BLLs or the flat spectrum radio quasars (FSRQs) using the SuperLearner, an ensemble method that includes several classification algorithms like logistic regression, support vector classifiers, Random Forest, Ranger Random Forest, multivariate adaptive regression spline (MARS), Bayesian regression, and extreme gradient boosting. We find that a SuperLearner model using MARS regression and Random Forest algorithms is 91.1 per cent accurate for kNN-imputed data and 91.2 per cent for MICE-imputed data. Furthermore, the kNN-imputed SuperLearner model predicts that 892 of the 1519 unclassified blazars are BLLs and 627 are FSRQs, while the MICE-imputed SuperLearner model predicts 890 BLLs and 629 FSRQs in the unclassified set. Thus, we can conclude that both imputation methods work efficiently and with high accuracy and that our methodology ushers the way for using SuperLearner as a novel classification method in the AGN community and, in general, in the astrophysics community.

Key words: methods: analytical – galaxies: active.

1 INTRODUCTION

Blazars are the extreme class of active galactic nuclei (AGNs) with jets oriented close to the observer's line of sight (Liodakis, Pavlidou & Angelakis 2017). They are traditionally composed of two main classes, BL Lacertae objects (BLLs) and flat spectrum radio quasars (FSRQs). The main distinction between these classes of objects is that BLLs often show no or very weak emission line spectra, whereas FSRQs typically show broad emission lines. The origin of the different blazar classes is still a matter of debate. However, it is often attributed to either a difference in accretion modes (e.g. Ghisellini et al. 2011) or observational biases (e.g. Padovani et al. 2019).

The orientation of the jets close to the line of sight causes extreme relativistic and projection effects (Liodakis et al. 2017; Liodakis et al. 2017, 2018), challenging the predominant classification scheme (e.g. Kharb, Lister & Cooper 2010). Blazars are the most numerous extragalactic γ-ray sources detected by the *Fermi* γ-ray space telescope (*Fermi*; Abdollahi et al. 2020). Out of the *Fermi* blazars, BLL dominates the population, accounting for more than 40 per cent of the sources. FSRQs account for only about 20 per cent, while blazars of uncertain type (BCUs) are about 35 per cent of all *Fermi* blazars (Ajello et al. 2022). Given this disparity, the γ-ray characteristics of BLLs and FSRQs may be distinct enough to be used as classification diagnostics, as opposed to the relative strength of atomic line spectra, as has been done historically. Human-based classification becomes highly inefficient, especially in multivariate data spaces. In this regard, machine learning (ML) is essential to detect patterns that otherwise would be unseen. We can broadly divide supervised ML into regression and classification. In the former, a value is estimated from the data, while in the latter, the class of an object is predicted based on a data set. Previous attempts using ML methods have focused on predicting blazar properties such as redshift (e.g. Dainotti et al. 2021; Gibson et al. 2022; Narendra et al. 2022), synchrotron peak (e.g. Glauch, Kerscher & Giommi 2022), and classification of *Fermi* BCUs and unidentified sources (e.g. Chiaro et al. 2016; Liodakis & Blinov 2019; Finke, Krämer & Manconi 2021; Coronado-Blázquez 2022). Liodakis & Blinov

★ E-mail: coopern@usmma.edu

Table 1. 4FGL (see also table 6 in Abdollahi et al. 2022).

Variable	Units	Description
Source name	–	Based on RA and Dec.: 4FGL JHHMM.m+DDMM.
Class	–	Class designation of likely associated source.
Association	–	Name of identified or likely associated source.
Energy Flux100	erg∗cm^{-2}∗s^{-1}	Energy flux in from 100 MeV to 100 GeV.
Highest energy	MeV	Energy of the highest-energy photon, association probability $P > 0.95$.
Variability index	–	Difference between flux per time interval and average flux.
Fractional variability	–	Fractional variability computed from fluxes each year.
Gaia G magnitude	–	*G*-band magnitudes from the *Gaia* survey.
$\nu_{\rm syn}$	s^{-1}	Synchrotron-peak frequency in observer frame.
$\nu F\nu_{\rm syn}$	erg∗cm^{-2}∗s^{-1}	Spectral energy distribution at synchrotron-peak frequency.
Spectrum type	–	The spectral type (PL, LP, and PLSuperExpCutoff).
PL Index	–	Photon index for the PL fit.
LP Index	–	Photon index at pivot energy for LP fit.
LP Beta	–	Curvature parameter for LP spectrum type.
LP SigCurv	σ units	Significance of the improvement between PL and LP fits.
LP Epeak	MeV	Peak energy in νF_ν estimated from the LP model.
PLEC IndexS	–	Photon index at pivot energy when fitting with PLSuperExpCutoff.
PLEC ExpfactorS	–	Spectral curve at pivot energy when fitting with PLSuperExpCutoff.
PLEC Exp Index	–	Exponential index when fitting with PLSuperExpCutoff.
PLEC SigCurv	σ units	Significance of the improvement for the PLSuperExpCutoff model to PL.
PLEC EPeak	MeV	Peak energy in νF_ν estimated for the PLSuperExpCutoff model.
Redshift	–	Redshift of identified or likely associated source.

(2019) combined infrared data and optical polarization observations from the RoboPol survey (Ramaprakash et al. 2019; Blinov et al. 2021) in Random Forest and logistic regression networks to identify blazars in the unidentified *Fermi* uncertainty regions. They verified the candidate AGN proposed by Mandarakas et al. (2019) for 3FGL J0221.2+2518 and paved the way for future optopolarimetric surveys (e.g. Polar-Areas Stellar Imaging in Polarization High-Accuracy Experiment (PASIPHAE); Tassis et al. 2018) to find more γ-ray-emitting AGNs. Dainotti et al. (2021), for the first time, used SuperLearner, an ensemble ML method that leverages the advantages of several constituent ML methods, in the AGN community to derive the redshift of AGNs with an accuracy that is comparable to other methods shown in the literature (Brescia et al. 2013, 2019). In the following papers (e.g. Gibson et al. 2022; Narendra et al. 2022), we show the reliability of the Multivariate Imputation by Chained Equation (MICE; Van Buuren & Groothuis-Oudshoorn 2011) in solving the problems of AGN variables missing at random and how these confirm the results of the accuracy of the redshift inference. Considering that a larger sample with more complete variables is essential for the classification purposes in the current analysis, we adopted the methods of MICE and SuperLearner for the AGN classification.

This study uses regression to estimate the data missing in the fourth *Fermi* Catalog (4FGL; Abdollahi et al. 2020) and classification to distinguish between BLLs and FSRQs among the BCUs. This has already been attempted by other groups in different wavelengths (e.g. Kovačević et al. 2020; Chiaro, Kovacevic & La Mura 2021; Bhat & Malyshev 2022; Butter et al. 2022; Sahakyan, Vardanyan & Khachatryan 2023) using single models. Kovačević et al. (2020) leverage a single artificial neural network (ANN) for classifying 1329 BCUs into 801 BLLs and 406 FSRQs, with 122 remaining unclassified. Bhat & Malyshev (2022) use Random Forest and neural network models to classify unassociated *Fermi* catalogue objects into pulsars and AGNs. Kang et al. (2019) use support vector machines (SVMs), ANNs, and Random Forest, achieving 91.8–92.9 per cent accuracy and predict that the unclassified blazars consist of 724 BLLs and 332 FSRQs with 332 still remaining unclassified. Butter et al. (2022) use a Bayesian neural network for classifying *Fermi* blazars into BLLs and FSRQs and provide uncertainty estimates in their predictions. Chiaro et al. (2021) have improved the ANN used in Kovačević et al. (2020) and provided a larger classification result for the 4FGL data set. Therefore, in addition to using only data provided in the 4FGL, we attack this problem by building an ensemble of methods to find the optimal classification model.

The paper is organized as follows. In Section 2, we present the data used in the analysis. In Section 3, we lay down the methodology and build our models. In Section 4, we present our results, and in Section 5, we summarize our conclusions. We also included an appendix (see Appendix A) where we analyse classification for all AGN classes present, not just blazars.

2 DATA SAMPLE

As previously stated, our data sample is taken from 4FGL (Abdollahi et al. 2022). Data consist of 3770 observations of 22 variables. Table 1 contains a short description of the variables included in this study; a more extensive description of each variable is obtained from Ajello et al. (2020).

Data include 18 numerical variables related to γ-ray emission, variability, optical magnitude, and spectral characteristics. Table 2 contains the descriptive statistics for these measurements for the blazar classes, BLL, FSRQ, and BCU. The data also include two categorical variables, classification, and spectrum type. The data contain 2192 power-law (PL) spectrum sources, 1572 log-parabola (LP) spectrum sources, and 6 PL superexponential cut-off (PLSuperExpCutoff; see Ajello et al. 2022) sources. We created 'is_PL' as a categorical variable where blazars with a PL spectrum have is_PL of one, and those with other spectral distributions have is_PL of zero. Note that in this sample, we have missing data in 12 features: highest energy, $\nu_{\rm syn}$, $\nu F_{\nu\,\rm syn}$, *Gaia G* magnitude, the significance between the PL and the LP spectrum (LP SigCurv), the peak energy in the $\nu_{\rm syn}$, νF_ν spectrum (LP EPeak), the photon index

Table 2. Numerical data summary.

	Min.	First Qu.	Median	Mean	Third Qu.	Max.	NAs
Energy Flux100	3.708e−13	1.804e−12	3.128e−12	8.045e−12	6.314e−12	8.440e−10	
Highest energy	4.153	10.000	21.344	47.728	53.544	912.930	657
Variability index	1.51	12.38	19.96	237.52	49.52	84 340.76	
Fractional variability	0.0000	0.0000	0.3514	0.4076	0.6122	2.9100	
Gaia G magnitude	12.81	17.79	18.58	18.56	19.46	21.45	1691
$\nu_{\rm syn}$	0.000e+00	0.000e+00	1.000e+13	1.549e+16	1.995e+14	1.120e+19	633
$\nu F_{\nu\ \rm syn}$	0.000e+00	0.000e+00	7.600e−13	2.203e−12	1.990e−12	2.200e−10	633
PL Index	1.419	1.992	2.213	2.219	2.444	3.153	
LP Index	−0.048 98	1.890 16	2.134 92	2.134 09	2.390 93	3.204 18	
LP Beta	−0.116 16	0.041 01	0.094 94	0.147 02	0.176 60	1.000 00	
LP SigCurv	0.0000	0.8402	1.6431	2.2372	2.8842	42.6642	27
LP Epeak	0	187	871	160 352	5904	111 488 184	567
PLEC IndexS	−0.1597	1.8404	2.1006	2.0863	2.3454	3.2123	27
PLEC ExpfactorS	−0.101 75	0.043 55	0.106 34	0.227 94	0.222 49	5.154 43	27
PLEC Exp Index	0.2419	0.6667	0.6667	0.6662	0.6667	0.6667	27
PLEC SigCurv	0.0000	0.8771	1.7415	2.2767	2.8841	43.4651	27
PLEC EPeak	0	802	4691	26 337	16 196	4593 440	1526
Redshift	0.0000	0.2548	0.6330	0.8577	1.2740	6.4430	1901

at the pivot energy when fitting with the superexponential cut-off (PLEC IndexS), the spectral curve at pivot energy when fitting with the superexponential cut-off (PLEC ExpfactorS), the exponential index when fitting with the exponential cut-off (PLEC Exp Index), the significance of the improvement of the superexponential fit to the PL fit (PLEC SigCurv), the peak energy in the $\nu_{\rm syn}$, νF_ν spectrum estimated for the PLSuperExpCutoff model (PLEC EPeak), and redshift. Since peak flux contained 2210 missing entries, we dropped this feature from the analysis.

We show histograms to compare distributions of missing data for features with more than 27 missing entries (see Fig. 1).

3 METHODOLOGY

The methodology is mainly divided into two steps: first, we treat the missing data and, secondly, we evaluate the classification through the ensemble method of SuperLearner (Van der Laan, Polley & Hubbard 2007; Polley & Van der Laan 2010), including the accuracy of the individual ML models outside of the SuperLearner ensemble. The methodology is described below in the following sections.

3.1 Data handling

As this study focuses on the classification of blazars, we did not include non-blazars from the data set. The total blazars are 1457 BLLs, 794 FSRQs, and 1519 BCUs. We divided these data into two sets, 2251 classified blazars for the training and testing of the various models we use and 1519 BCUs for evaluation. However, each set contains missing data.

The set of classified blazars (BLLs and FSRQs) contains 927 missing PLEC EPeaks, 638 missing *Gaia G* magnitudes, 607 missing redshifts, 301 missing LP EPeaks, 215 missing highest energies, 201 missing for both $\nu_{\rm syn}$ and $\nu F\nu_{\rm syn}$, and 1 missing each for LP SigCurv, PLEC IndexS, PLEC ExpfactorS, PLEC Exp Index, and PLEC SigCurv.

The set of unclassified blazars (BCUs) contains 1294 missing redshifts, 1053 missing *Gaia G* magnitudes, 599 missing PLEC EPeaks, 442 missing highest energies, 432 missing for both $\nu_{\rm syn}$ and $\nu F\nu_{\rm syn}$, 266 missing LP EPeaks, and 26 each for LP SigCurv, PLEC IndexS, PLEC ExpfactorS, PLEC Exp Index, and PLEC SigCurv. We show the missing entry compared to the full set of variables in Fig. 2, which can be read as follows: for example, for the variable LP EPeak, 567 AGNs have missing entries. The number 567 is mentioned at the bottom of the figure corresponding to LP EPeak. Along the rows of the figure, for example, row 3, it reads as 206 AGNs having missing data in two variables. These two variables are *Gaia G* magnitude and redshift. We discuss our methods for handling these missing data below.

3.2 Missing data

Here, we have applied two methods for the imputation of the missing data. These are as follows:

(i) **Logistic regression:** To form a basis of comparison for the effect of the missing data, we have trained a logistic regression (logit) model by not using the measurements with missing data. Logistic regression calculates the log-odds of a binary outcome, the response variable, by one or more predictor variables. Log-odds are defined by

$$\ln\left(\frac{p}{1-p}\right), \tag{1}$$

where p is the probability (0 to 1) of the binary outcome. Here, the binary outcome is a BLL or an FSRQ, modelled by a variable that is 1 for BLLs and 0 for FSRQs. In the case of multiple predictor variables, not all will statistically impact the outcome, and thus we perform a back-elimination to remove such variables.

Back-elimination begins with all variables included in the calculation, and the model contains a significance coefficient, p-value (P), for each variable. In statistics, a $P = 0.05$ is the conventional threshold for statistical significance. We used the same convention here. First, we identify the variable with the largest p-value (weakest predictor) and remove it if $P > 0.05$. Then, the model is retrained, and the process is repeated, removing the consecutive weakest predictor with the $P > 0.05$. This process is repeated until only variables with p-values smaller than 0.05 (95 per cent certainty that the coefficient is different from zero) remain.

Regarding the logistic regression models, we have treated the missing data in two ways, by dropping sources (rows) that contain missing data and by dropping the 12 features (columns) with missing data

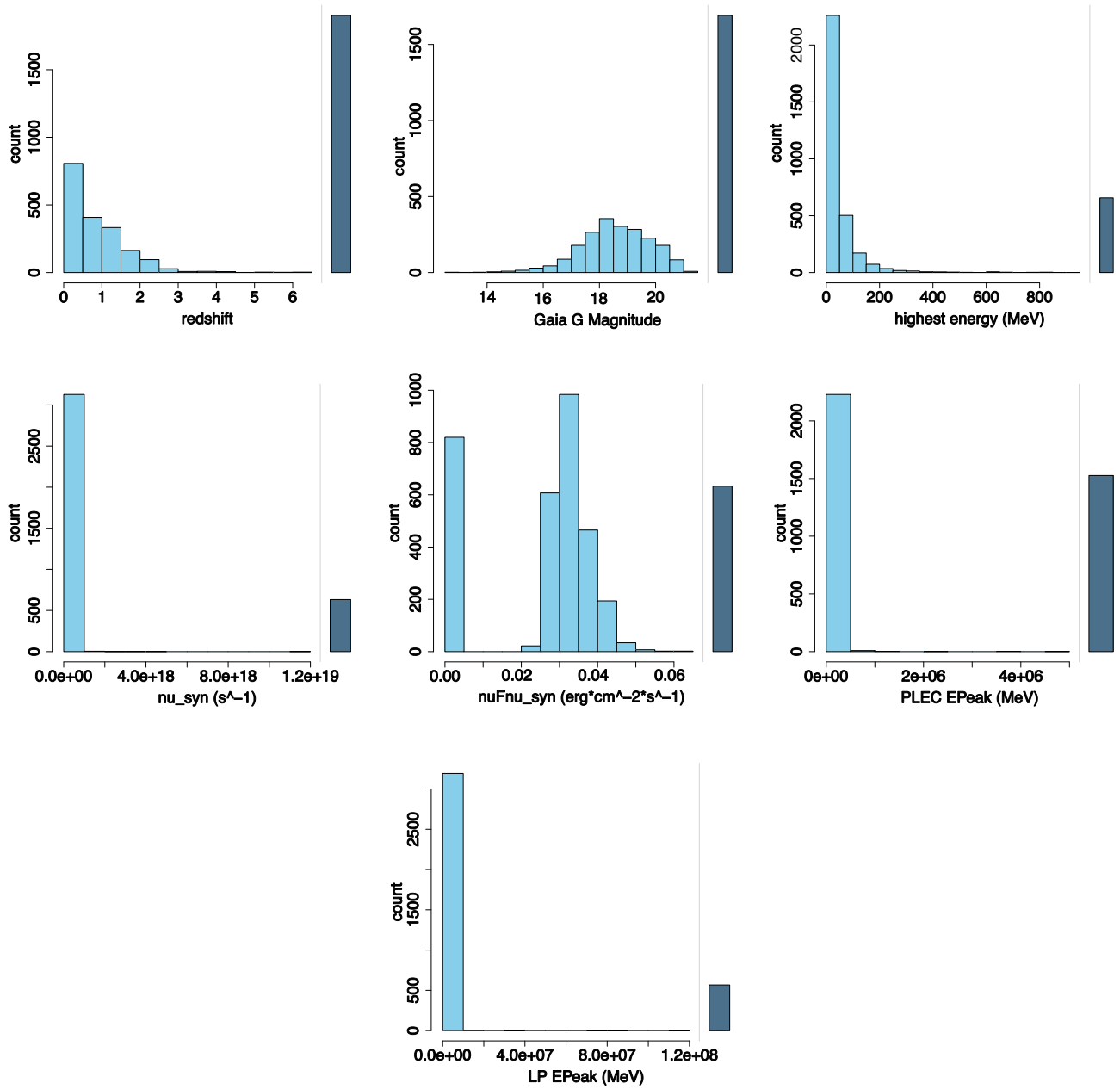

Figure 1. Top left: Redshift distribution of the blazar set, including the scale of the missing data. Top middle: *Gaia G* magnitude distribution of the blazar set, including the scale of the missing data. Top right: Highest energy (MeV) distribution of the blazar set, including the scale of the missing data. Middle left: $\nu_{\rm syn}$ ($\rm s^{-1}$) distribution of the blazar set, including the scale of the missing data. Middle: $\nu F\nu_{\rm syn}$ ($\rm erg{*}cm^{-2}{*}s^{-1}$) distribution of the blazar set, under eighth root transformation, including the scale of the missing data. Middle right: PLEC EPeak (MeV) distribution of the blazar set, including the scale of the missing data. Bottom: LP EPeak (MeV) distribution of the blazar set, including the scale of the missing data.

(see Fig. 2). There are 288 blazars out of 2251 that have complete data. In the sample where rows with missing data are dropped, there are 160 BLLs and 116 FSRQs, for an imbalance ratio (IR) of 1.38. The complete set of identified blazars contains 1457 BLLs and 794 FSRQs for an IR of 1.84. Dropping the columns does not affect the IR. In our logistic regression analysis, all features remain significant for the set of 288 classified blazars with complete data. Only 12 BCUs have complete data. The complete features that remained significant after back-elimination are the variability index, fractional variability, spectrum type quantified as 1 for PL and 0 for not PL, and the PL Index.

(ii) **MICE and kNN:** We estimated the missing data using imputation. We used two imputation methods, *k*-nearest neighbours (kNN; Fix & Hodges 1952) and MICE (Van Buuren & Groothuis-Oudshoorn 2011; Luken, Padhy & Wang 2021; Gibson et al. 2022). We use MICE since we have already successfully used it in previous analysis (Gibson et al. 2022), and then for comparison, we used the logistic regression, but we could have also used other models. Although we could try different approaches, this is beyond the scope of the current analysis. MICE uses a regression method on the present data to estimate a value for the missing data. Both are regression methods that calculate a numerical value based on the data

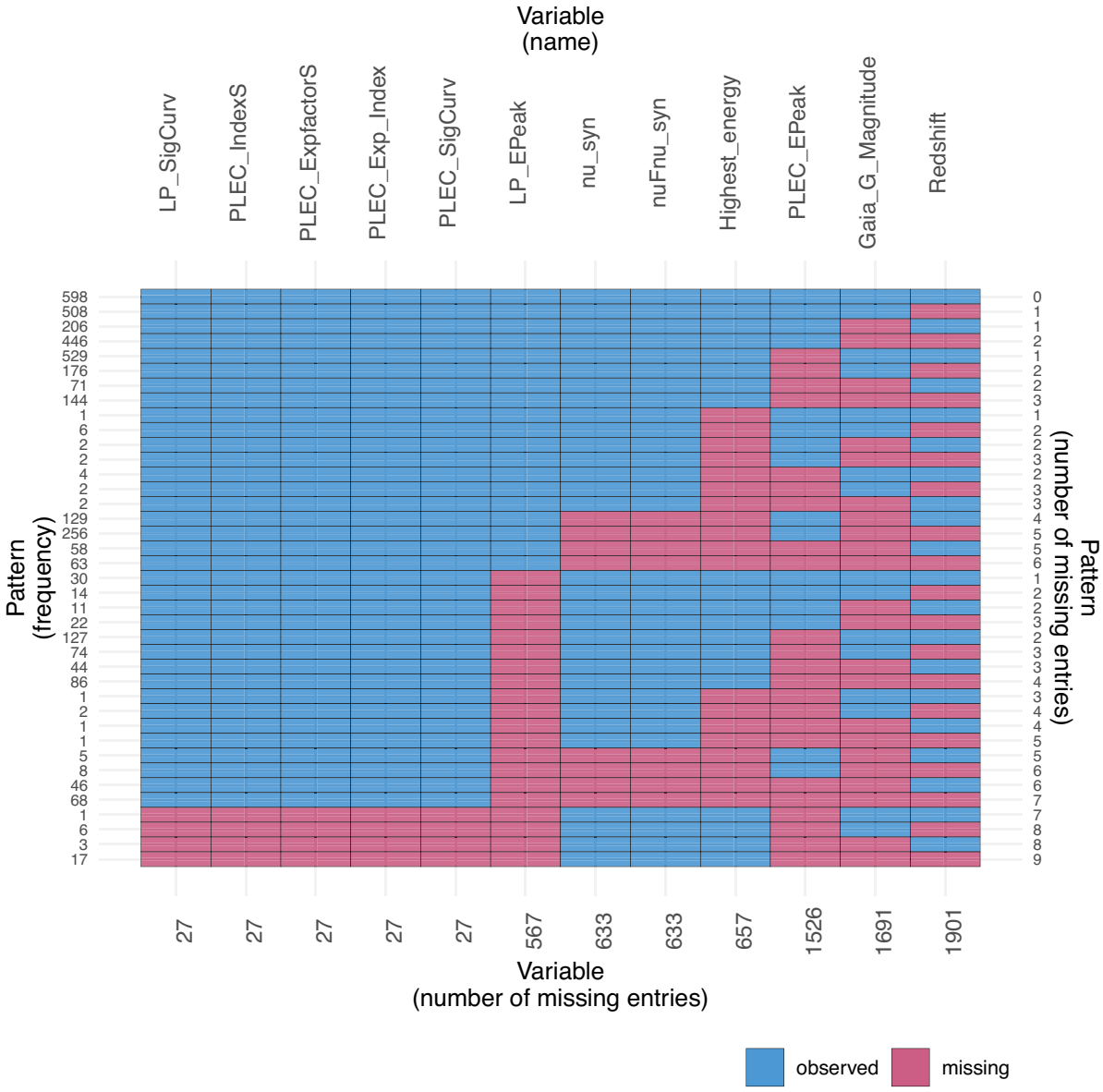

Figure 2. This figure shows all the missing data. The blue boxes show the complete data, and the magenta boxes show the missing data. The variable names are presented at the top of the figure, while the numbers at the bottom of the figure represent the number of missing entries for each variable. The number on the right shows the number of variables with missing entries, and the corresponding number on the left shows the number of AGNs that have missing entries in those variables.

provided. kNN estimates a value by finding the k number of closest observations with complete data to the observations with missing data. The Visualization and Imputation of Missing Values (VIM) library provided the kNN imputation function for this study. The VIM kNN algorithm uses the Gower (1971) coefficient of similarity, which allows the measurement of the similarity of items with mixed numeric and non-numeric data. The default k-value for kNN is equal to 5. To replace a missing variable for a given object, VIM kNN identifies its five nearest neighbours according to the Gower similarity coefficient. It then averages the measurement of interest in the five neighbours and assigns the average to the missing value in the incomplete observation.

The MICE algorithm uses a range of statistical techniques to impute an incomplete column (the target column) by generating 'plausible' synthetic values given to other columns in the data. In this work, we selected the Classification and Regression Tree ('CART') option to predict unknown values. The MICE function in R allows the generation of multiple imputations of a given value, which reflects uncertainty in the data (Van Buuren & Groothuis-Oudshoorn 2011; see Fig. 3). The $\nu F\nu_{\rm syn}$ feature required transformation to converge on a solution in MICE. We discuss this below in Section 3.3.

3.3 Data transformations

It is always advisable to proceed with the data transformation to improve the predictive power of ML. Indeed, in our case, when performing the MICE imputation, it could not converge on a solution for the $\nu F\nu_{\rm syn}$ feature, which has a bimodal distribution with one mode at zero and one mode distributed at around an order of magnitude of 10^{-12}. A logarithmic transformation is difficult to

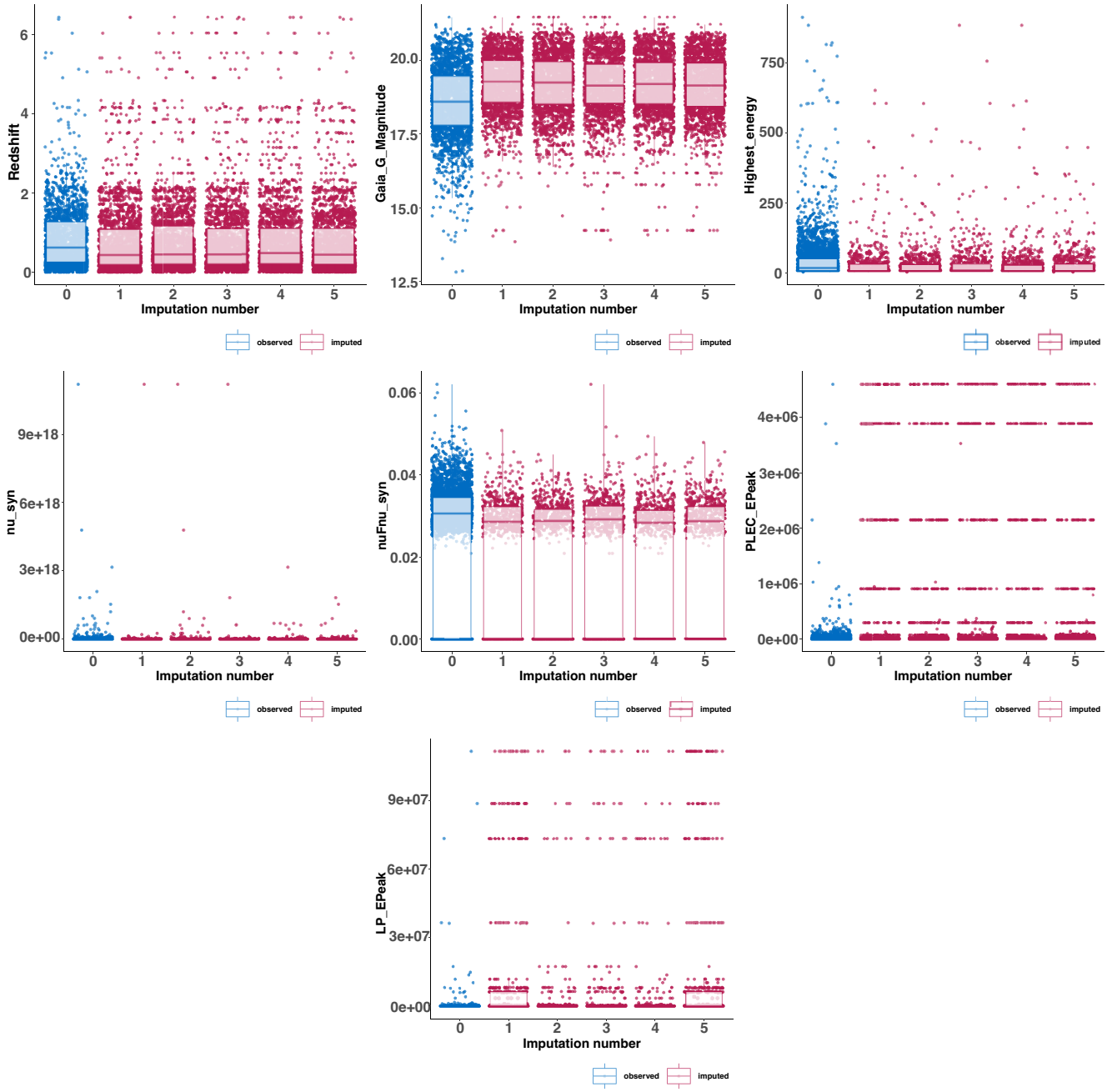

Figure 3. Distribution plots of observed data and the five MICE imputation streams. The box and whisker plots superimposed over the data show a box that represents the centre 50 per cent of the distribution with a 75-percentile upper boundary, the middle bar is the median, and the low boundary is the 25-percentile, the vertical lines above and below the box show the upper and lower 25-precentiles, respectively.

perform due to the mode at zero. Instead, we performed an eighth root transformation. This transformation shifted the second mode to the order of 10^{-2}, and separated the two modes enough to allow MICE to converge on a solution.

Furthermore, we performed a z-score normalization. Certain ML algorithms, such as logistic regression, are sensitive to the order of magnitude and may misidentify variables as being non-statistically significant if the units' order of magnitude is much less than other variables in the set. For example, Energy Flux100 feature has an order of magnitude of 10^{-12}, whereas $\nu_{\rm syn}$ has an order of magnitude of 10^{13}, as seen in Table 2. ML algorithms may ignore Energy Flux100 based on its low order of magnitude due to the units used and not the underlying physics. The equation for z-score normalization, which mitigates the effect of the order of magnitude differences, is

$$m_z = \frac{m - \bar{m}}{\sigma_m}, \tag{2}$$

where m is the measurement, $\bar{m}$ is the mean of the measurements, and σ_m is the standard deviation. In effect, the z-score is a fractional measure of how many standard deviations an individual measurement varies from the mean. The z-scoring was performed for all numerical data after imputation occurred.

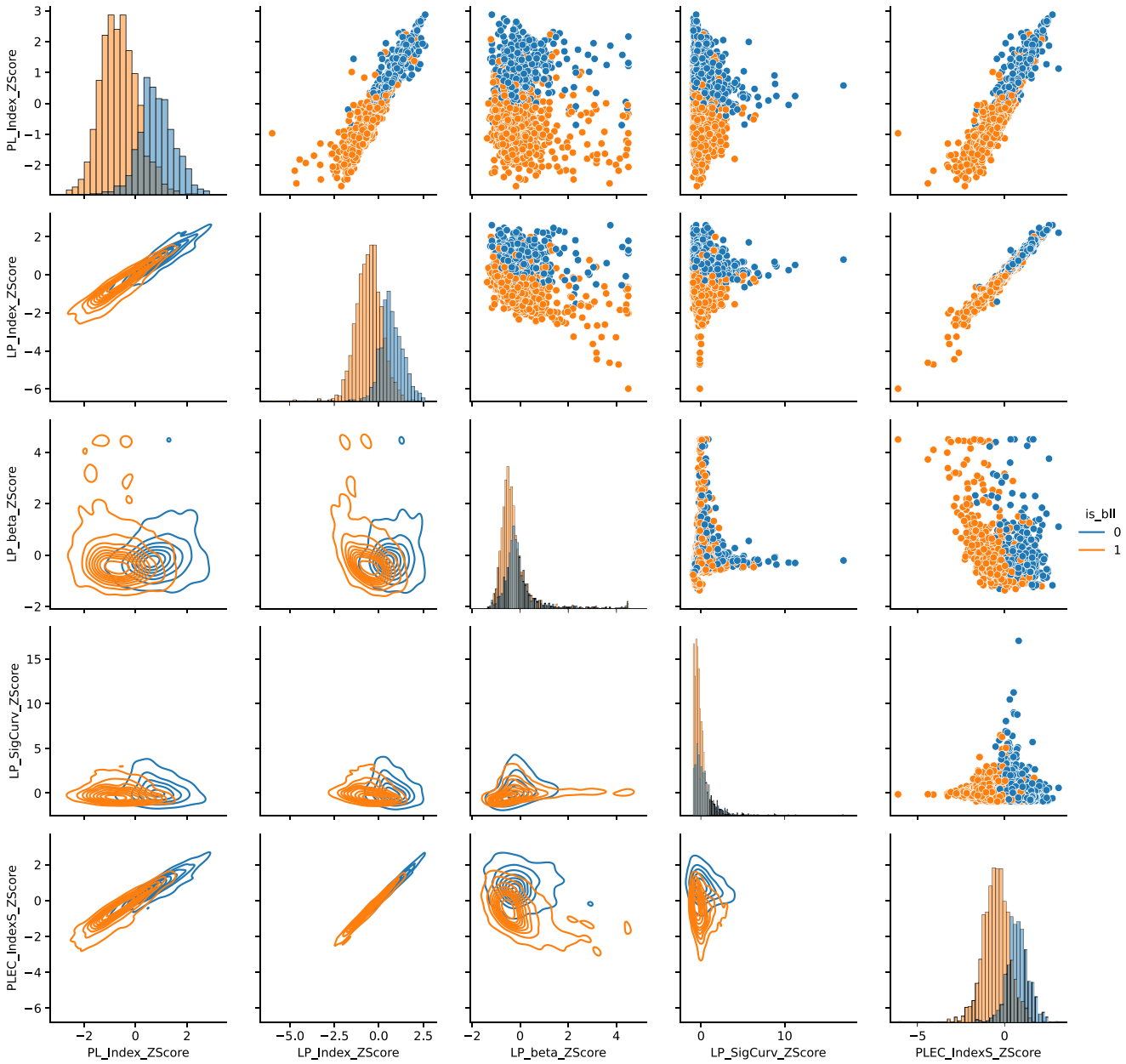

Figure 4. Pair grid plot of spectrum variables, MICE. The orange are the BLLs and the blue are the FSRQs.

Figs 4 and 5 show a pair grid plot. Along the diagonal are histograms of the z-score normalized data. Above the diagonal are scatter plots. Below the diagonal are kernel density plots, the contours showing the percentile of data within its boundary. All data are colour-coded such that BLLs are orange and FSRQs are blue.

3.4 Machine learning models

Upon imputing the missing data using kNN and MICE, we built several classification models to be used in the SuperLearner ensemble algorithm. As with handling the missing data, we used back-elimination to eliminate variables that had no statistically significant effect on the outcome of the logistic regression models. After logistic regression, the models we used are SVMs, Random Forest, Ranger Random Forest, neural networks, Enhanced Adaptive Regression Through Hinges (EARTH) regression, Bayesian regression, and extreme gradient boosting (XGB). We tested each algorithm separately for both kNN- and MICE-imputed data. Ultimately, SVMs, neural networks, and Bayesian regression were eliminated from the final MICE-imputed and kNN-imputed SuperLearner models. Below is a short description of the models included in the final MICE and kNN ensemble SuperLearner models. We use the default parameters of the R functions unless otherwise noted; in Bayesian, Ranger, and EARTH regressions, we had good performance using the defaults. Our concern is that overtuning of the parameters will bias the models to the training sets, and hinder performance when the models are eventually used against new data in future data releases.

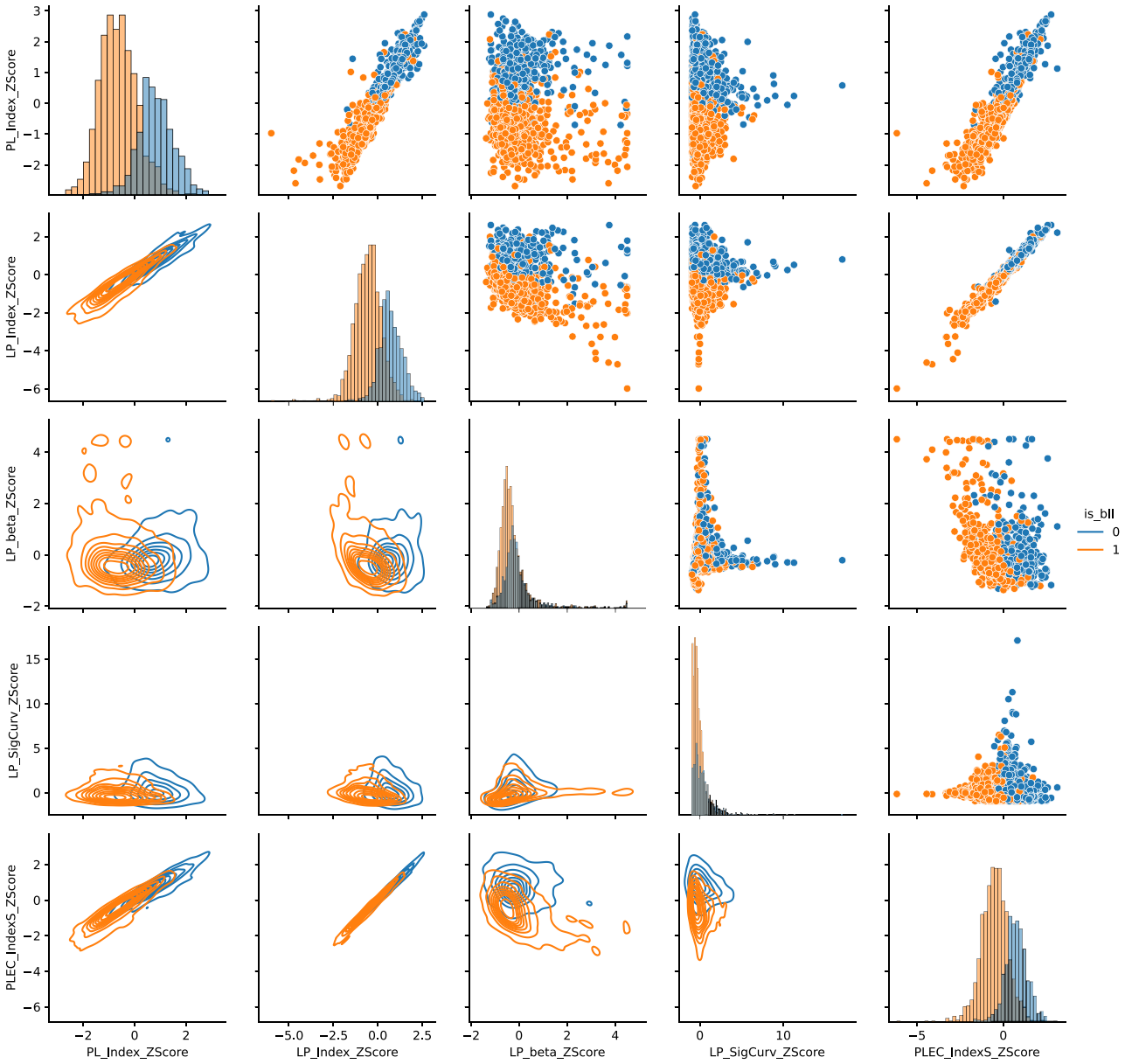

Figure 5. Pair grid plot of spectrum variables, kNN. The orange are the BLLs, while the blue are the FSRQs.

3.4.1 Parametric methods: EARTH

EARTH is based on the multivariate adaptive regression spline (MARS; Friedman & Roosen 1995) method. The EARTH algorithm allows for better modelling of predictor interaction and non-linearity in the data compared to the linear model. EARTH fits a sum or product of hinges, which are partwise linear fits of the data that are combined so that the sum-of-squares residual error is minimized with each added term. In this study, we set the family variable, i.e. the kernel to binomial.

3.4.2 Decision tree methods: Random Forest, extreme gradient boosting, and Ranger

Random Forest is a supervised learning algorithm of the decision tree class. Decision trees measure conditional probabilities of the predictor variables with respect to the response variable and then sort the response variables according to the values of the individual predictor variables. This algorithm tends to overfit the predictor variables (Ho 1995). Random Forest is a method of compensating for the overfitting of individual decision trees by randomly selecting subsets of data from the training data and creating many decision trees. The final prediction for the response variable is based on the majority outcome of the decision trees for a classification problem. In this study we used the random Forest function in the RANDOM FOREST library. We set the max trees to 500 and the maximum number of nodes per tree to 10.

The Ranger algorithm is nearly the same as a standard Random Forest, except that it uses extremely randomized trees (ERTs). ERTs share the top-down procedure as Random Forest, except that the data cuts are randomly generated, assuming a uniform distribution of the predictor variables in the training set. The cut chosen is the one that

minimizes prediction error (Geurts, Ernst & Wehenkel 2006). The ranger function in R requires that the classification variable be set to true.

XGB is also a decision tree class algorithm. We use the XGBOOST library for R. Gradient boosting creates an ensemble of decision trees as weak learners. These weak learners are then built into strong learners using an iterative gradient descent method to minimize a loss function such as root mean squared error (RMSE). XGB differs from other gradient boosting algorithms by using the Newton−Ralphson method instead of a gradient decent method. The Newton−Ralphson method is a root-finding algorithm used to minimize the loss function. In this study, we set booster to 'gbtree', eta to 0.001, max depth to 5, gamma to 3, subsample to 0.75, 'colsample_bytree' to 1, objective to 'multi:softprob', 'eval_metric' to 'mlogloss', and 'num_class' to 2.

3.4.3 *Neural network*

In ML, ANNs are models that mimic biological neural networks. ANNs are a network similar to those in the mathematical discipline of Graph Theory in that the network is composed of nodes and edges. Nodes are the objects in the network, and edges are the connections between the objects. As ANNs are trained, statistical weights are assigned to edges between nodes that, based on input data, determine which pathway an observation takes through the network. These pathways will then determine the outcome. For an ANN classifier, the outcome is the predicted class of the observation (Jain, Mao & Mohiuddin 1996). In this study, using the NEURALNET library in R, we set hidden node layers to 3 and 1, the activation function to logistic, and 'linear.output' to false.

3.4.4 *Support vector machine*

SVMs are a type of supervised ML model that can be used for either classification or regression. SVMs can classify multiple classes because they find boundaries between classes within the *N*-dimensional space created by the data, where *N* is the number of features in the data. The SVMs form these boundaries, called support vectors, using kernel functions such as linear kernels, polynomial kernels, radial basis function kernels, or sigmoid kernels. In this study, we used the SVM function from the E1071 library, type was set to 'C-classification', the radial kernel was used, the cost was set to 10, and the scale to false (Cortes & Vapnik 1995).

3.4.5 *Bayesian regression*

Bayesian regression is a conditional model that uses Bayes Theorem to predict a feature, such as the numerical class of an AGN, based on a linear combination of the other features of the data set. The regression coefficients are calculated via the posterior probability distribution of the features. In this study, we used the bayesglm function in the ARM library. We use the normal linear model, which assumes that the dependent feature is a function of a Gaussian distribution of the independent features (Raftery 1996). The Gaussian distribution is well suited for our study since we use *z*-score normalization, which measures the standard deviation from the mean for each feature.

3.4.6 *SuperLearner*

SuperLearner (Van der Laan et al. 2007; Dainotti et al. 2021; Gibson et al. 2022; Narendra et al. 2022) is an ensemble method that leverages on the advantages of several ML models and weights each constituent model according to its predictive power. It constructs models using a 10-fold cross-validation (CV). SuperLearner also calculates a set of normalized coefficients. The user defines the models used for SuperLearner's input and determines the final prediction by minimizing RMSE (Polley & Van der Laan 2010). Note that the final model is trained on 9 folds of the CV and the 10th fold is used on the test set, and this CV is internal to the SuperLearner function. The CV we perform in individual model testing is a separate procedure. The CV normalized coefficients are higher for models with lower RMSE. The condition required by the SuperLearner is that the sum of the coefficients (A_i), which corresponds to the predictive power of the constituent models, should be 1. We use these coefficients to determine which models will be included in the final model as explained below in Section 3.5, and the family variable was set to binomial.

3.5 Model building procedure

The process for each algorithm is the same. The data have been separated into randomly selected training sets and test sets. We sorted 80 per cent of the data into the training set and 20 per cent into the test set. The models are trained on the training set. Randomly selected training/test sets are useful to prevent overfitting the model.

When overfitting occurs, the model exhibits high accuracy in classifying the training set but much lower accuracy when the model is used on new data. To address this issue, we built each model, regardless of the algorithm, using CV (see Section 3.4.6). In CV, multiple training–test sets are created to test whether the model is overfitting to the training data and to quantify the effect of random selection between the training and test sets on the final model. Then, it is used to classify the objects in the test set. If a model is not overfitted, it should have high accuracy when classifying the test set. We use 100 nested 5-fold CV. For each of the 100 iterations, the training/test sets are selected randomly, and the model's accuracy is measured on the test set. The histogram for 100 accuracy measurements is provided in Fig. 6.

We also created the confusion matrices (Altman & Bland 1994) for the performance of each model when validating the model using the test set. A confusion matrix is an $n \times n$ matrix, where n is the number of classes that map the actual classes on the column space and the predicted classes on the row space. The diagonals of the matrix are true predictions, i.e. BLLs predicted to be BLLs, and FSRQs predicted to be FSRQs. In Fig. 7, BLLs are mapped as 1 by the 'is_bll' variable, the upper right are BLLs falsely identified as FSRQs, and the bottom left is for FSRQs falsely identified as BLLs. The confusion matrix is useful for identifying whether a model has a tendency for false positives (FPs), FSRQs (is_bll = 0) identified as BLLs ('is_bll' = 1), or false negatives (FNs), BLLs identified as FSRQs. From these confusion matrices, we calculate accuracy, which is the ratio of correct classifications from the test set to the size of the test set.

Similar to our procedure for logistic regression, we used the back-elimination process for each algorithm used in SuperLearner to select the optimal model for the final SuperLearner ensemble. Back-elimination was based on the normalized coefficient that SuperLearner put on each algorithm, with those <5 per cent on average being eliminated. The following algorithms were not eliminated from the final model: EARTH regression, XGB, Random Forest, and Ranger Random Forest. Both MICE- and kNN-imputed SuperLearner models used StepAIC for predictor variable selection, which is an independent process from the model selection.

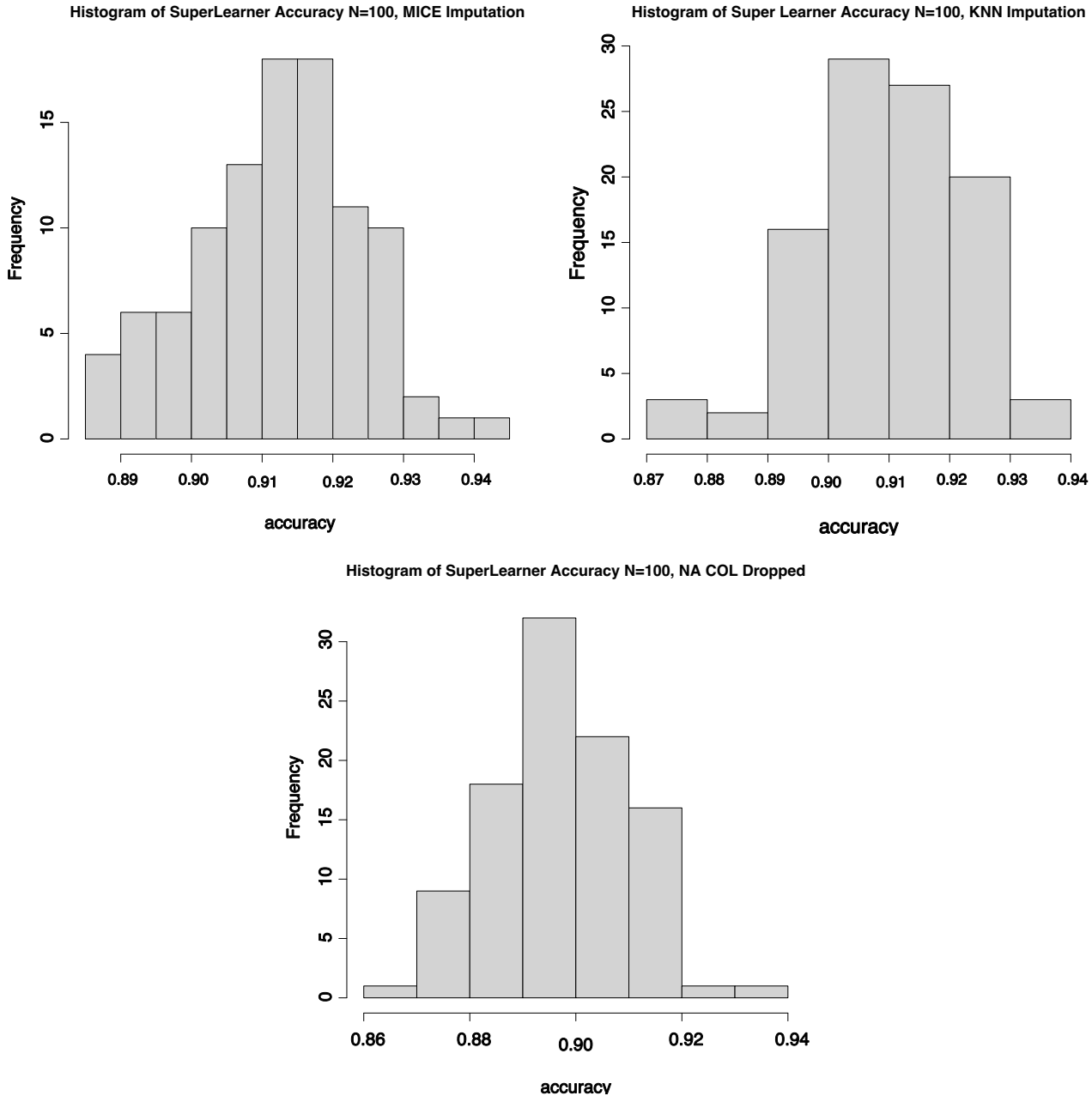

Figure 6. Histograms of SuperLearner model accuracy. Top left: SuperLearner with MICE-imputed data. Top right: SuperLearner with kNN-imputed data. Bottom: SuperLearner with NA columns dropped.

StepAIC works similarly to back-elimination in that non-significant predictor variables are removed from the model. However, the statistic used is the Akaike information criterion (*AIC*; Sakamoto, Ishiguro & Kitagawa 1986):

$$AIC = 2(k - \ln(\hat{L})), \tag{3}$$

where k is the number of parameters in the model and $\hat{L}$ is the maximized value from the likelihood function. *AIC* is a measurement of how well an individual model performs relative to the set of models used (Van der Laan et al. 2007).

4 RESULTS

SuperLearner performed well with both MICE and kNN imputation of missing data, with 91.2 and 91.1 per cent average accuracy, respectively, on the test set. As mentioned in Section 3.1, there is an IR of 1.84. The SuperLearner model with MICE-imputed data predicted that of the 1519 BCUs, 890 are BLLs and 629 are FSRQs for an IR of 1.41. The SuperLearner model with kNN-imputed data predicted 892 BLLs and 627 FSRQs for an IR of 1.42. Despite the imputation method, SuperLearner predicts a slightly lower ratio of BLLs in the unclassified blazars than in the classified set. The accuracies of the other models are shown in Table 3.

The logistic regression model using kNN-imputed data had an accuracy of 90.3 per cent. The statistically significant variables were

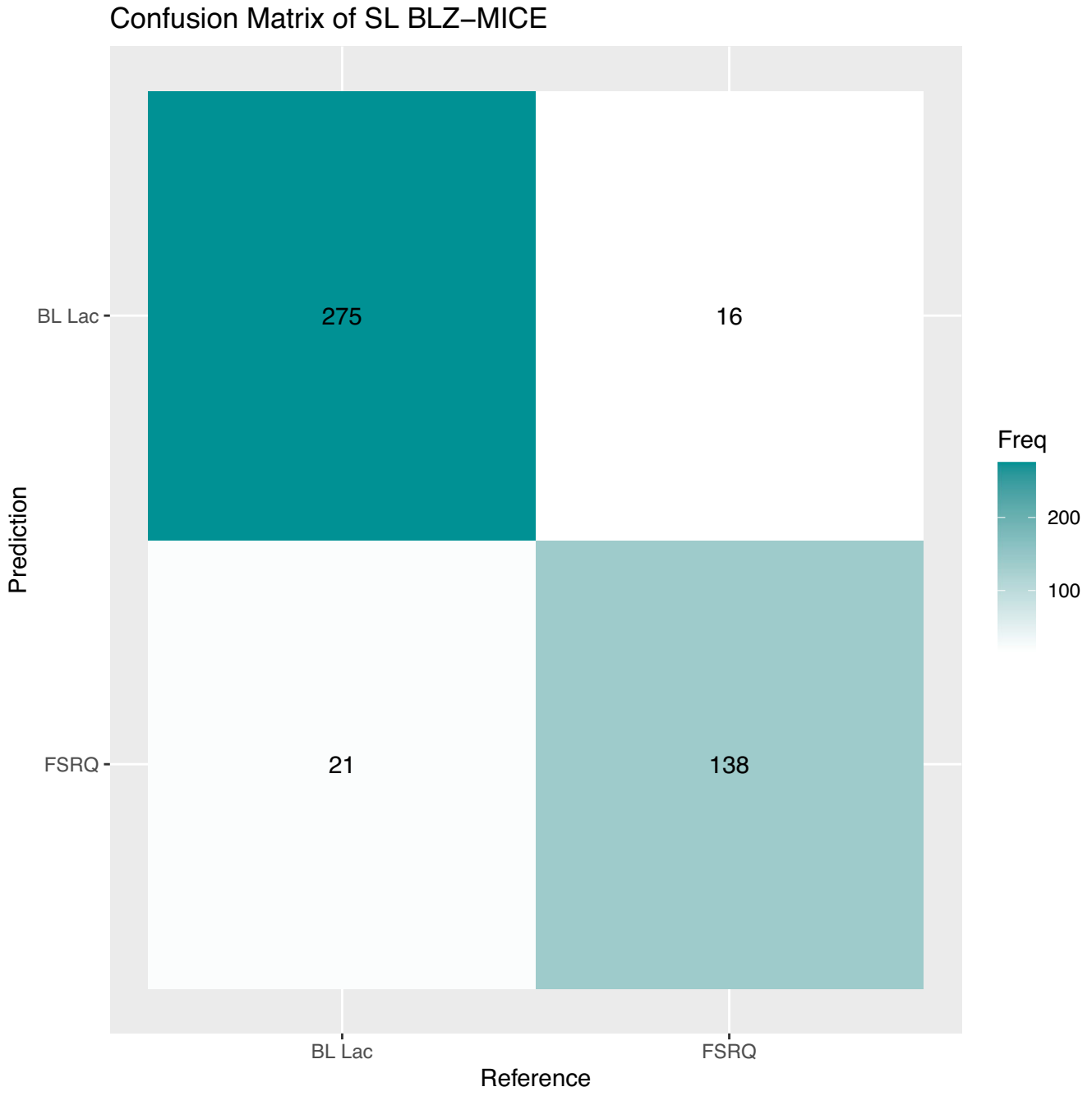

Figure 7. Confusion matrix plot of an individual SuperLearner model with an accuracy of 91.8 per cent, a sensitivity of 89.6 per cent, a specificity of 92.9 per cent, and a balanced accuracy of 91.3 per cent. Reference indicates the class of the object in the 4FGL DR3 catalogue. Prediction is the predicted class of the SuperLearner model.

fractional variability, is_PL, PL Index, and redshift. The logistic regression model trained on kNN-imputed data classified the BCUs into 900 BLLs and 619 FSRQs for an IR of 1.45. The logistic regression model with MICE-imputed data has an accuracy of 90.3 per cent. The statistically significant variables are fractional variability, *Gaia G* magnitude, PL Index, LP SigCurv, and redshift. When the MICE logistic regression model is used to predict the classification of the BCUs, it results in 875 BLLs and 644 FSRQs for an IR of 1.36. The logistic regression model with missing data removed by feature has an accuracy of 89.8 per cent. The statistically significant variables are variability index, fractional variability, is_PL, and PL Index. When the logistic regression model is used to predict the classification of BCUs, it results in 896 BLLS and 623 FSRQs for an IR of 1.44.The logistic regression model missing data removed by observation has an accuracy of 89.2 per cent. The IR was 1.38. There are only 12 BCUs with complete data, and the model predicts 10 FSRQs and 2 BLLs. Note that the weak spectra of BLLs bias this result.

As stated above, the data have an IR of 1.84. In dealing with imbalanced data, Löpez et al. (2013) note that problems with imbalanced data often do not arise from the IR itself but from other issues with the data, such as overlapping class separability, which

Table 3. Model accuracy, 5-fold CV, 10*x*nested for neural networks, and 100*x*nested for other models.

Model	MICE accuracy (per cent)	kNN accuracy (per cent)
Logistic regression	90.3	90.3
Random Forest	91.1	90.8
SVM	90.0	90.3
XGB	91.2	91.4
Bayesian regression	90.4	90.3
EARTH	89.9	90.2
Ranger	91.0	91.1
Neural networks	89.7	90.0
SuperLearner	91.2	91.1

Table 4. Classification statistics, 5-fold CV, and 100*x*nested SuperLearner MICE imputation.

Statistic	Score
Sensitivity	0.87
Specificity	0.94
PPV	0.89
NPV	0.93
F1	0.88
Prevalence	0.35
Detection rate	0.31
Detection prevalence	0.35
Balanced accuracy	0.90

we can see in Figs 4 and 5. Their analysis for overlapping class separability starts with a data set with an IR of 9. In Johnson & Khoshgoftaar (2019), a meta-analysis of studies on ML with high IR, the smallest ratio of IR in the meta-analysis is 2. Taken together, it seems that an IR of 1.84 is below the threshold for using special techniques other than using other statistics than simple accuracy, which we will discuss below.

The accuracy of each model averaged over 100 times nested 5-fold CV is shown in Table 3; the histograms of the SuperLearner models are shown in Fig. 6. Please note that the MICE method tends to underperform kNN, albeit not beyond fluctuation due to the random selection of training and test sets. Imputing missing data outperforms excluding missing data in the logistic regression and SuperLearner models. From Table 3, we can see that the best-performing models are Ranger, XGB, and SuperLearner. Overall, there is not a large difference among the models, although the SuperLearner, as expected, has one of the highest predictions.

We also have created a SuperLearner model using the data where columns with missing data are dropped, as these data are performed the same as in the logistic regression model. When the missing data are not accounted for, SuperLearner model has an accuracy of 89.8 per cent under nested 5-fold CV (see bottom panel of Fig. 6). This is only slightly less accurate than the imputed data models, which are 91.2 and 91.1 per cent for MICE and kNN, respectively. When used to predict the class of the BCUs, SuperLearner predicts 890 BLLs and 629 FSRQs for a ratio of 1.41. Regardless of the data treatment, SuperLearner predicts a lower ratio of BLLs to FSRQs than the data in the training and test sets.

Table 4 shows the classification statistics for the MICE-imputed SuperLearner classifier averaged over the 100 nested CVs. The confusion matrix function in R interprets FSRQs as positive and BLLs as negative. *Sensitivity* is also called the true positive (TP) rate; this means that the ratio of correctly classified FSRQs to total FSRQs in the test set is 87 per cent. Specificity is also called the true negative (TN) rate. This is the rate of correctly classified BLLs to total BLLs in the test set, 94 per cent. Balanced accuracy is the arithmetic average of these two scores, 90 per cent. Note that the accuracy score reported in Table 3 is the number of correct classifications over the sample size. The two scores do not agree because BLLs outnumber FSRQs by 1457 to 794. Given this class imbalance, the balanced accuracy is the more reliable of the two accuracy measures. Positive predictive value (PPV) is the ratio of correctly identified FSRQs to the sum of correctly identified FSRQs and BLLs falsely identified as FSRQs, 89 per cent, and negative predictive value (NPV) is the ratio of correctly identified BLLs to the sum of correctly identified BLLs to FSRQs falsely identified as FSRQs, 93 per cent. The F1 score is the harmonic mean of the *PPV* and *sensitivity*. The F1 score is calculated by

$$F_1 = \frac{2 * PPV * Sensitivity}{PPV + Sensitivity}. \quad (4)$$

In statistics, harmonic means, such as the F1 score, are used to calculate the average rate; here, F1 is 88 per cent.

Prevalence is the number of FSRQs in the sample over the total sample size, 35 per cent. The detection rate is the number of correctly predicted FSRQs over the total sample size, 31 per cent. The detection prevalence is the number of all predicted FSRQs over the sample size, 35 per cent.

5 SUMMARY AND CONCLUSIONS

Our aim is to classify AGNs using several classification techniques embedded in the SuperLearner, such as EARTH regression, XGBOOST, Random Forest, and Ranger Random Forest. To allow the maximum number of observations with all the features, we use MICE. We also use kNN imputation as a comparison to MICE. We here summarize below our major findings, which are relevant both for the classification methodology and from a physical point of view:

(i) MICE imputation is statistically tied with kNN imputation with every algorithm.

(ii) Using imputation of missing data outperforms dropping data with missing variables along either features (columns) or observations (rows).

(iii) Spectral parameters, such as PL Index and LP SigCurv, and variability parameters, such as fractional variability, are significant in logistic regression modelling.

(iv) 100 GeV γ-ray flux is not significant in any logistic regression model.

(v) Redshift's significance in classification may support the notion that BLLs and FSRQs, and by extension, Fanaroff–Riley type I and type II radio galaxies (RDGs), may be evolutionary steps in radio-loud AGNs.

(vi) Although XGB has slightly higher accuracy, note that this could be due to statistical fluctuation, we expect SuperLearner to perform better on new data, as ensemble methods are less likely to overtrain on the training data.

Redshift's significance in classification paves the way for a better understanding of this class of objects. In addition, we use the SuperLearner for the first time in the realm of classification. We point out that SuperLearner has already been successfully used in the realm of regression (Dainotti et al. 2021; Gibson et al. 2022; Narendra et al. 2022). The use of SuperLearner here opens a new perspective on astrostatistics analysis. It allows the application of

powerful ML methods that are still unknown or rarely used in the astronomy community. When new data are available and the new AGN catalogue is released, we will be able to further support this scenario or highlight new subtle features which otherwise would remain buried in the classification.

ACKNOWLEDGEMENTS

This research was supported by the Visibility and Mobility module of the Jagiellonian University (Grant number: WSPR.WSDNSP.2.5.2022.5) and the NAWA STER Mobility Grant (Number: PPI/STE/2020/1/00029/U/00001). AN is grateful to the Exploratory Research Grant for the financial support for the visit of AN to the Division of Science at the National Astronomical Observatory of Japan.

DATA AVAILABILITY

The data used in this paper are taken from the Fourth *Fermi* Large Area Telescope AGN Catalog (Abdollahi et al. 2022). Post-processed data are available upon request.

REFERENCES

Abdollahi S. et al., 2020, ApJS, 247, 33
Abdollahi S. et al., 2022, ApJS, 260, 53
Ajello M. et al., 2020, ApJ, 892, 105
Ajello M. et al., 2022, ApJS, 263, 24
Altman D. G., Bland J. M., 1994, BMJ, 308, 1552
Bhat A., Malyshev D., 2022, A&A, 660, A87
Blinov D. et al., 2021, MNRAS, 501, 3715
Brescia M., Cavuoti S., D'Abrusco R., Longo G., Mercurio A., 2013, ApJ, 772, 140
Brescia M., Salvato M., Cavuoti S., Ananna T. T., Riccio G., LaMassa S. M., Urry C. M., Longo G., 2019, MNRAS, 489, 663
Butter A., Finke T., Keil F., Krämer M., Manconi S., 2022, J. Cosmol. Astropart. Phys., 2022, 023
Chiaro G., Kovacevic M., La Mura G., 2021, J. High Energy Astrophys., 29, 40
Chiaro G., Salvetti D., La Mura G., Giroletti M., Thompson D. J., Bastieri D., 2016, MNRAS, 462, 3180
Coronado-Blázquez J., 2022, MNRAS, 515, 1807
Cortes C., Vapnik V., 1995, Mach. Learn., 20, 273
Dainotti M. G. et al., 2021, ApJ, 920, 118
Finke T., Krämer M., Manconi S., 2021, MNRAS, 507, 4061
Fix E., Hodges J. L., 1952, Report A193008, Nonparametric Discrimination: Small Sample Performance. USAF School of Aviation Medicine, Randolph Field, Texas
Friedman J. H., Roosen C. B., 1995, Stat. Methods Med. Res., 4, 197
Geurts P., Ernst D., Wehenkel L., 2006, Mach. Learn., 63, 3
Ghisellini G., Tavecchio F., Foschini L., Ghirlanda G., 2011, MNRAS, 414, 2674
Gibson S. J., Narendra A., Dainotti M. G., Bogdan M., Pollo A., Poliszczuk A., Rinaldi E., Liodakis I., 2022, Frontiers Astron. Space Sci., 9, 836215
Glauch T., Kerscher T., Giommi P., 2022, Astron. Comput., 41, 100646
Gower J. C., 1971, Biometrics, 27, 857
Ho T. K., 1995, in Kavanaugh M., Storms P., eds, Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1: ICDAR '95. IEEE Computer Society, USA, p. 278
Jain A., Mao J., Mohiuddin K., 1996, Computer, 29, 31
Johnson J. M., Khoshgoftaar T. M., 2019, J. Big Data, 6, 27
Kang S.-J., Li E., Ou W., Zhu K., Fan J.-H., Wu Q., Yin Y., 2019, ApJ, 887, 134
Kharb P., Lister M. L., Cooper N. J., 2010, ApJ, 710, 764
Kovačević M., Chiaro G., Cutini S., Tosti G., 2020, MNRAS, 493, 1926
Liodakis I. et al., 2017, MNRAS, 466, 4625
Liodakis I., Blinov D., 2019, MNRAS, 486, 3415
Liodakis I., Hovatta T., Huppenkothen D., Kiehlmann S., Max-Moerbeck W., Readhead A. C. S., 2018, ApJ, 866, 137
Liodakis I., Pavlidou V., Angelakis E., 2017, MNRAS, 465, 180
Löpez V., Fernández A., García S., Palade V., Herrera F., 2013, Inf. Sci., 250, 113
Luken K. J., Padhy R., Wang X. R., 2021, preprint (arXiv:2111.13806)
Mandarakas N. et al., 2019, A&A, 623, A61
Narendra A., Gibson S. J., Dainotti M. G., Bogdan M., Pollo A., Liodakis I., Poliszczuk A., Rinaldi E., 2022, ApJS, 259, 55
Padovani P., Oikonomou F., Petropoulou M., Giommi P., Resconi E., 2019, MNRAS, 484, L104
Peduzzi P., Concato J., Kemper E., Holford T. R., Feinstein A. R., 1996, J. Clin. Epidemiology, 49, 1373
Polley E. C., Van der Laan M. J., 2010, U.C. Berkeley Division of Biostatistics Working Paper Series Vol. 266. Available at: https://biostats.bepress.com/ucbbiostat/paper266
Raftery A. E., 1996, Biometrika, 83, 251
Ramaprakash A. N. et al., 2019, MNRAS, 485, 2355
Sahakyan N., Vardanyan V., Khachatryan M., 2023, MNRAS, 519, 3000
Sakamoto Y., Ishiguro M., Kitagawa G., 1986, Akaike Information Criterion Statistics, Vol. 1, Mathematics and its Applications. Boston: D. Reidel, Dordrecht, p. 290
Tassis K. et al., 2018, preprint (arXiv:1810.05652)
Van Buuren S., Groothuis-Oudshoorn K., 2011, J. Stat. Softw., 45, 1
Van der Laan M. J., Polley E. C., Hubbard A. E., 2007, Stat. Appl. Genetics Mol. Biol., 6, 1

APPENDIX: CLASSIFICATION OF ALL AGN TYPES

This appendix shows SuperLearner's performance when all identified AGN types are included. The AGN classes included in Data Release 3 (DR3) are BLLs, FSRQs, and BCUs, as above, as well as 45 RDGs, 9 AGNs with no further subtype, 8 narrow-line Seyfert 1s (NLSY1s), 5 compact steep spectrums (CSSs), 2 Seyferts (SEYs), and 2 steep spectrum radio quasars (SSRQs).

For logistic regression classifiers, the accepted practice is to have 10 observations for 1 feature (Peduzzi et al. 1996). We use 18 features in this analysis. Under the accepted practice, the minimum number of observations needed for a reliable result is 180. Only BLLs and FSRQs meet that threshold, with BCUs being withheld for classification. However, the accepted 10:1 ratio was established using solely logistic regression. SuperLearner is an ensemble method. Below, we test whether SuperLearner can perform well with fewer than 10 observations per feature.

Of the non-blazar classes, only the 45 RDGs have a count high enough to be represented in both the training and test sets. Under binomial probability distribution, the expected number of RDGs in the training set is 36, and there is a 0.004 per cent probability that no RDGs appear in the test set. However, for the 9 AGNs, the expected number of AGNs in the training set is 7.2, and the probability of no AGN in the test set is 13.4 per cent.

SuperLearner requires that the training and test sets have the same number of classes. SuperLearner also only supports binomial classification, that is, classification between two classes. To meet these two requirements, we created three binary classifiers. As with the blazar analysis, we created 'is_bll' with 1 representing a BLL and 0 representing another class. Additionally, we created 'is_fsrq' where 1 is for FSRQs and 0 for non-FSRQs, and finally we created 'is_rdg' where 1 is for RDGs and 0 for another class. The other classes, AGNs, NLSY1s, CSSs, SEYs, and SSRQs, with too few

Table A1. Classification statistics, 5-fold CV, and 100*x*nested SuperLearner MICE imputation – all AGNs.

	Sensitivity	Specificity	*PPV*	NPV	F1	Prevalence	Detection rate	Detection prevalence	Balanced accuracy
Class: nbagn	0.09	0.97	0.04	0.99	NaN	0.01	0.00	0.03	0.53
Class: bll	0.93	0.84	0.91	0.87	0.92	0.63	0.58	0.64	0.88
Class: fsrq	0.85	0.94	0.87	0.92	0.86	0.34	0.29	0.33	0.89
Class: rdg	0.06	1.00	NaN	0.98	NaN	0.02	0.00	0.00	0.53

observations to reliably appear in both the training and test sets, were reclassified as 'NBAGN' for non-blazar AGNs.

During the CV loop, we trained three instances of SuperLearner, one to classify BLLs, one to classify FSRQs, and one to classify RDGs. Each class is represented by a binary vector: (1, 0, 0) for BLLs, (0, 1, 0) for FSRQs, and (0, 0, 1) for RDGs, and all other combinations are classified as NBAGNs. The resultant classification statistics are in Table A1. Note that with small numbers of RDGs and NBAGNs, if low numbers of these objects were contained in the test sets, *PPV*, NPV, and prevalence and, by extension, F1 value can have division by zero, resulting in a not a number (NaN) value. For NBAGNs, due to the low count, sensitivity or *PPV* for an individual instance in the CV can be zero, resulting in an F1 value of NaN since F1 is proportional to $1/(PPV + Sensitivity)$. For RDGs, the *PPV* is NaN resulting in an F1 value of NaN. We see in Table A1 that the sensitivity for RDGs is 6 per cent, and the sensitivity for NBAGNs is 9 per cent. Specificity is 97 per cent for NBAGNs and 100 per cent up to two significant figures for RDGs. So, while the model does not accurately predict RDGs or NBAGNs, they tend not to misclassify other objects as RDGs or NBAGNs. Note that the performance of the models for BLLs and FSRQs is comparable to the binomial classifier (see Table 4). Specifically, since the binomial classifier treated FSRQs' positive cases comparing Table 4 to the FSRQ row in Table A1, we see that FSRQs' multinomial sensitivity is 87 per cent, specificity is 94 per cent, *PPV* is 87 per cent, NPV is 92 per cent, F1 is 86 per cent, prevalence is 34 per cent, detection rate is 29 per cent, detection prevalence is 33 per cent, and balanced accuracy is 88 per cent. These stats are all approximately 1–2 per cent less than the binomial case, and the combined number of RDGs and NBAGNs represents 3.1 per cent of the total sample. This suggests that the RDGs and NBAGNs misclassified as blazar types bring these average statistics down.

Fig. A1 shows the confusion matrix corresponding to one model. In this model, NBAGN sensitivity is 0 per cent with 0 TPs and 5 FNs, and specificity is 97 per cent with 403 TNs and 14 FPs; RDG sensitivity is 20 per cent with 1 TP and 4 FNs, and specificity is 100 per cent with 402 TNs and 0 FPs; BLLs have a sensitivity of 93 per cent with 268 TPs and 20 FNs, and a specificity of 81 per cent with 135 TNs and 31 FPs; and FSRQs have a sensitivity of 81 per cent with 134 TPs and 32 FNs, and a specificity of 94 per cent with 269 TNs and 16 FPs. It is also noteworthy how this uneven class count affects sensitivity and specificity with the more numerous class, BLLs being weighted towards higher sensitivity and lower specificity, and FSRQs, RDGs, and NBAGNs being weighted towards higher specificity and lower sensitivity. It is also worth noting that this model demonstrates the 100 per cent specificity of the RDGs, which means that the model does not misidentify other classes as RDGs.

We conclude in this appendix that these data support the findings of Peduzzi et al. (1996) regarding observation to features ratio. To accurately and positively predict other AGN categories, more observations of RDGs and the types that make NBAGNs are needed, although the 45 observations of RDGs do seem to be enough to prevent FPs of that class.

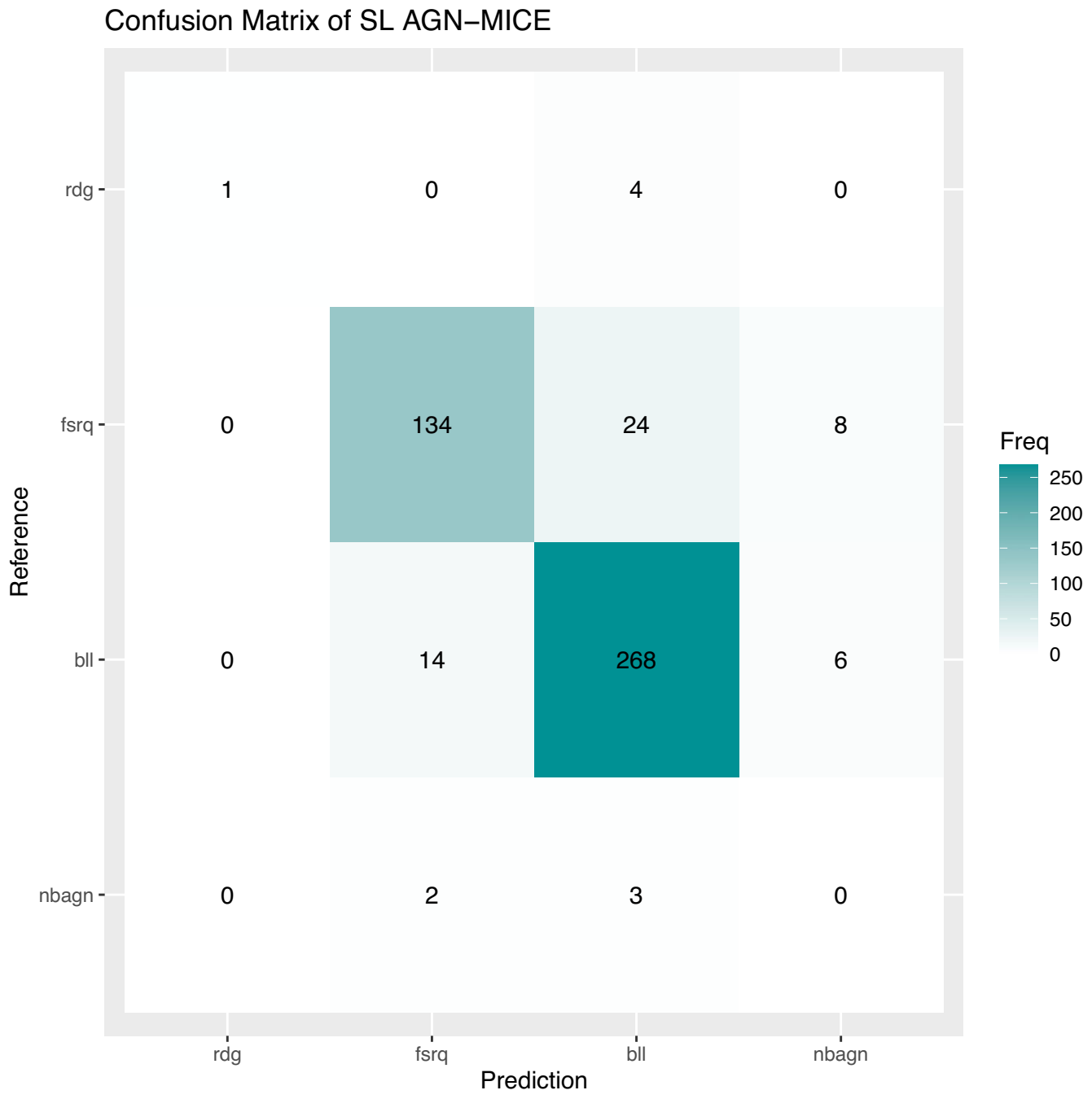

Figure A1. Confusion matrix plot of an individual SuperLearner model with an accuracy of 86.9 per cent, a sensitivity of 89.6 per cent, a specificity of 92.9 per cent, and a balanced accuracy of 91.3 per cent. The reference indicates the class of the object in the 4FGL DR3 catalogue. Prediction is the predicted class of the SuperLearner model.

This paper has been typeset from a TEX/LATEX file prepared by the author.

Published by Oxford University