Article

Exploring New Redshift Indicators for Radio-Powerful AGN

Rodrigo Carvajal [1,2,*], Israel Matute [1,2], José Afonso [1,2], Stergios Amarantidis [1,2], Davi Barbosa [1,2], Pedro Cunha [3,4] and Andrew Humphrey [3]

1   Instituto de Astrofísica e Ciências do Espaço, Universidade de Lisboa, OAL, Tapada da Ajuda, PT1349-018 Lisbon, Portugal; imatute@oal.ul.pt (I.M.); jafonso@iastro.pt (J.A.); samarant@oal.ul.pt (S.A.); dbarbosa@oal.ul.pt (D.B.)
2   Departamento de Física, Faculdade de Ciências, Universidade de Lisboa, Edifício C8, Campo Grande, PT1749-016 Lisbon, Portugal
3   Instituto de Astrofísica e Ciências do Espaço, Universidade do Porto, CAUP, Rua das Estrelas, PT4150-762 Porto, Portugal; pedro.cunha@astro.up.pt (P.C.); andrew.humphrey@astro.up.pt (A.H.)
4   Departamento de Física e Astronomia, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre 687, PT4169-007 Porto, Portugal
*   Correspondence: rcarvajal@oal.ul.pt

Abstract: Active Galactic Nuclei (AGN) are relevant sources of radiation that might have helped reionising the Universe during its early epochs. The super-massive black holes (SMBHs) they host helped accreting material and emitting large amounts of energy into the medium. Recent studies have shown that, for epochs earlier than $z \sim 5$, the number density of SMBHs is on the order of few hundreds per square degree. Latest observations place this value below 300 SMBHs at $z \gtrsim 6$ for the full sky. To overcome this gap, it is necessary to detect large numbers of sources at the earliest epochs. Given the large areas needed to detect such quantities, using traditional redshift determination techniques—spectroscopic and photometric redshift—is no longer an efficient task. Machine Learning (ML) might help obtaining precise redshift for large samples in a fraction of the time used by other methods. We have developed and implemented an ML model which can predict redshift values for WISE-detected AGN in the HETDEX Spring Field. We obtained a median prediction error of $\sigma_z^N = 1.48 \times (z_{\mathrm{Predicted}} - z_{\mathrm{True}})/(1 + z_{\mathrm{True}}) = 0.1162$ and an outlier fraction of $\eta = 11.58\%$ at $(z_{\mathrm{Predicted}} - z_{\mathrm{True}})/(1 + z_{\mathrm{True}}) > 0.15$, in line with previous applications of ML to AGN. We also applied the model to data from the Stripe 82 area obtaining a prediction error of $\sigma_z^N = 0.2501$.

Keywords: Active Galactic Nuclei; radio galaxies; redshift determination; multiwavelength catalogues; Machine Learning

PACS: 98.54.Cm; 98.54.Gr; 98.62.Py; 95.75.Pq; 95.80.+p

1. Introduction

Super-Massive Black Holes (SMBHs) might be ubiquitous to all galaxies above a certain mass. Understanding their true role in the shaping of galaxies will require a more precise census of the nature, growth, and evolution of SMBHs—in the so-called Active Galactic Nuclei (AGN) phases—, as well as a more detailed characterisation of the internal (secular) and external (environment) processes at work within the host [1].

Radio selection has been traditionally a prime wavelength for the detection of AGN activity. Between 10–20% of AGN have strong radio emission, in many cases in the form of jets, that can overshadow the radio emission associated to star-forming regions, mostly due to super-novae [2]. Radio selection efficiency though seems to decrease towards the Epoch of Reionisation (EoR. $z > 6$, e.g., Refs. [3–6]). Simulations (e.g., Refs. [7–9]) predict that the distribution of AGN and Radio Galaxies (RG) along redshift can lead to the detection of a few hundreds of objects per square degree at the EoR as the with deep observations planned

for future observatories, e.g., SKA, µJy sensitivity levels [10]. These expectations collide with the most recent compilations (see, for instance, Refs. [11,12]), which show that only $\sim$300 sources have been confirmed to exist at $z > 6$ over most of the sky. Environmental (CMB) and intrinsic (QSO versus radio mode accretion) conditions might be responsible for the lower rate of radio powerful sources at $z > 5$ but, selection criteria, might also be playing a role [13]. Nevertheless, current radio instruments and recently completed surveys [14–17] have allowed detection of larger numbers of RG (e.g., Refs. [18–20]) that could be used to better understand the origin and evolution of radio emission in AGN.

To place radio AGN in the proper cosmological context and derive their intrinsic properties, and given the time constrains imposed for the compilation of significant spectroscopic samples, alternative estimates for redshift need to be used. Template-based photometric redshifts have proven to be an efficient alternative by trading precision for sky coverage. The sizes of the new catalogues though, with tens to hundred of millions of sources, imply a significant—and ever increasing—investment in computational time. These issues raise the need for additional approaches which might be able to obtain the redshift information for a large number of astrophysical sources with enough precision and within a reasonable amount of time.

The tremendous increase of computing power over the last decades has allowed the application of evolved statistical methods in the analysis of large and complex datasets. Using previously-fed data, it is possible to predict, with relevant confidence, the behaviour new data will have. This is what has been called Machine Learning (ML). In Astrophysics, ML has been used in a wide range of subjects (in AGN and other types of sources), such as redshift determination (e.g., Refs. [21,22]), morphological classification (e.g., Refs. [23–27]), source selection and classification (e.g., Refs. [28–31]), image and spectral reconstruction (e.g., Ref. [32]), and more [33,34]. Despite its range of applications, ML has received some criticism related to the interpretability of the derived models, e.g., most ML models cannot provide coefficients that allow to create an analytical expression for example [35]. This implies that it may not be straightforward to understand the exact role that the measured properties have into the prediction a model might make.

Recent work has been done to improve interpretability. Feature importance [36] can be derived, mostly for Tree-Based models, i.e., models that use decision trees to classify or predict properties. In this scenario, a feature with a high importance will be, in general, in the higher levels of the decision trees used for the modelling. A different method for assessing the impact of features is that of Shapley Values [37]. Opposite to feature importance, Shapley values, which have been defined in the context of Game Theory to determine the contribution of a player in a cooperative game, can help in understanding how the features impact each individual prediction. A more thorough description on how Shapley values work can be seen in Molnar [38].

Astronomical data is very heterogeneous in its current form, with small areas of the sky covered extensively at all wavelengths and with high sensitivity but also larger areas with sparser multiwavelength coverage. Therefore, the homogeneous and deep multiwavelength coverage required for the most accurate models can only be achieved over a few to tens of degrees. The derived models in these fields could then be applied first to present surveys with less extensive and deep multiwavelength coverage (e.g., LoTSS, Stripe82, RACS, MIGHTEE, etc.) and then also to the upcoming all-sky surveys, e.g., SKA, LSST, eROSITA, etc., delivering observations with comparable depth and multiwavelength coverage as current small fields.

In this work, we describe an ML model aiming to predict the redshift for AGN based on the multiwavelength properties of the HETDEX Spring Field with the minimum amount of data preparation possible. The model will be then tested in data from the SDSS Stripe 82 Field where multiwavelength coverage and depth vary with respect to the HETDEX Spring Field. This approach would test the validity of the derived ML model based on a field into other fields with slightly different spectral coverage.

The structure of this article is as follows. In Section 2, we present the used data, its preparation for ML training and describe the selection of models and the metrics used for assess their results. In Section 3, the results of model training and validation are shown, as well as the predictions over the Stripe 82 field. We present the discussion of our results in Section 4. Finally, in Section 5, we summarise our work.

2. Materials and Methods

2.1. Data

We have selected all the detections listed on the CatWISE2020 catalogue (CW, [39]) that are located in the HETDEX Spring Field and that have been covered by the LOFAR DR1 measurements [17] (see Figure 1). The CatWISE2020 catalogue has measurements in the WISE bands W1 and W2 (at 3.4 µm and 4.6 µm, respectively), with a 90% completeness depth at W1 = 17.7 mag and W2 = 17.5 mag. LOFAR observations cover an area of 424 deg^2, with a median sensitivity of 71 µJy/beam and a 6″ resolution. In that way, we have obtained 6,729,647 detected sources.
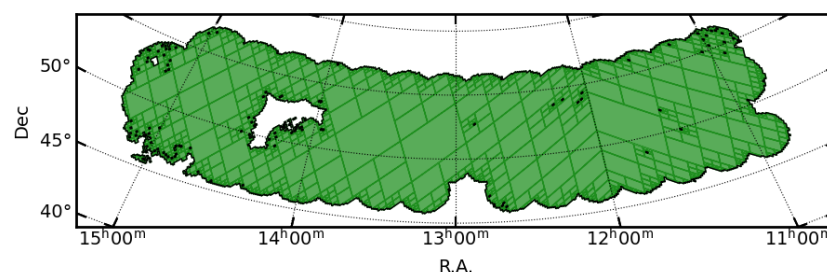

Figure 1. Area of the HETDEX Spring Field covered by the LOFAR DR1 measurements. Figure prepared, in part, using the Python package `MOCPy` [40].

The sources have been then cross-matched with other catalogues in different wavelengths using a search radius of 5″. We have selected surveys with large sky coverages, such as: VLASS (3 GHz) [16], LoTSS-DR1 (150 MHz) [17], Pan-STARRS DR1 [41,41], GALEX AIS GR6+7 [42], GMRT 150 MHz all-sky [43], 4XMM-DR9 [44], 2MASS All-Sky [45,46], and All-WISE (AW [47]). The 20 selected photometric bands are listed in Table 1. To homogenise photometric measurements, we converted all fluxes and magnitudes to AB magnitudes.

Table 1. Photometric bands included in the dataset.

Survey/Instrument	Bands	Survey/Instrument	Bands
CatWISE2020	W1, W2	VLASS	3.0 GHz
AllWISE	W1, W2, W3, W4	GALEX	FUV, NUV
Pan-STARRS	g, r, i, z, y	2MASS	J, H, K
LOFAR	150 MHz	XMM-NEWTON	0.2–12 keV
GMRT	150 MHz		

We then selected the sources that could be linked to the emission of an AGN. Thus, we cross-matched our catalogue with the Million Quasar Catalog[1] (MQC, v7.2, [48]). It lists published type-I QSOs/AGN, quasar candidates, type-II object and blazars along with the best available redshift values for each of them, i.e., spectroscopic or photometric redshifts. For the HETDEX Spring Field, 32,365 objects have been identified, in different surveys, as AGN. That means that 0.48% of the detected CatWISE2020 sources have been identified as AGN, and close to 8% (2674) of them are considered as QSO candidates. From the identified AGN in our sample, 26,520 sources are listed in the Sloan Sky Digital Survey Quasar Catalogue DR16 (SDSS-Q DR16 [49]), implying that the mean properties of the objects studied in this work are driven by the behaviour of SDSS QSOs. In Figure 2, we show the distribution of the sources in the WISE colour-colour diagram and the histogram of available redshifts.
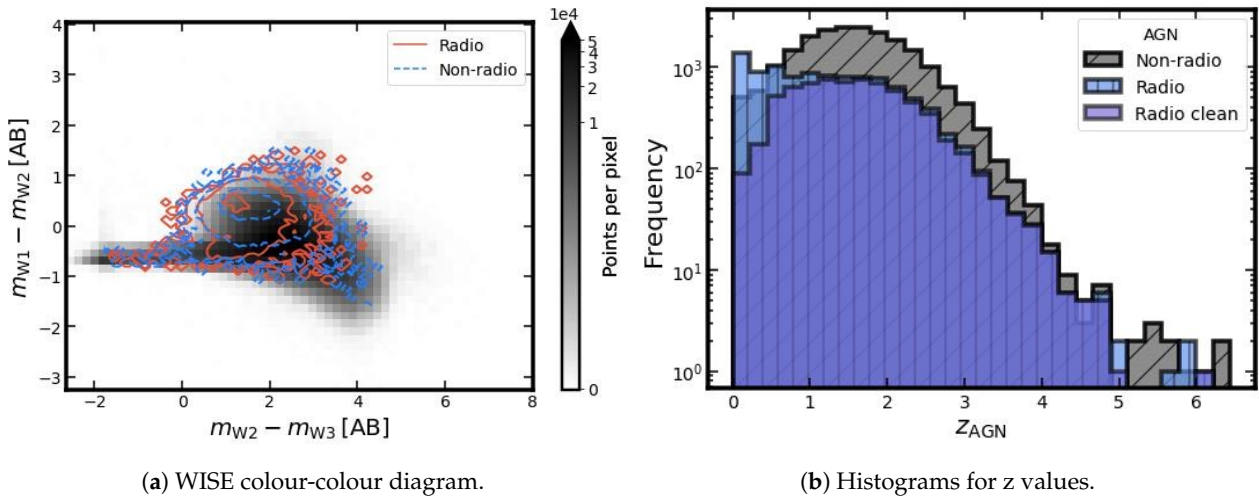
(**a**) WISE colour-colour diagram.

(**b**) Histograms for z values.

Figure 2. Characterisation plots for the Active Galactic Nuclei (AGN) sources in the HETDEX Spring Field. (**a**) W1 - W2, W2 - W3 colour-colour diagram. Grey background represents the full CatWISE2020 sample, with darker areas showing higher number of sources following the colour bar. Red, solid contours show density levels for radio-detected AGN. Blue, dashed contours indicate density levels for AGN without a counterpart on the radio surveys used in this work (i.e., without radio detection). For both contour plots, the lines show the levels with 1, 10, 100, and 1000 sources in each pixel. (**b**) Histograms for the redshift values of sources labelled as AGN. Grey, hatched histogram shows the distribution of redshifts for AGN without radio detections. Redshifts for all radio-detected AGN are presented by the blue, vertically-hatched histogram. Confirmed AGN without high host influence (see main text) that show a measurement on the surveys used in this work, are presented in purple, clean histogram.

One important feature to note in Figure 2b is the distribution of radio-detected sources (i.e., sources which show a counterpart on either LOFAR, GMRT, or VLASS). It does not follow the same trends as non-radio AGN, i.e., there is not a peak around $z = 2$ and the number of sources increases towards $z = 0$. This behaviour appears from the inclusion of all AGN listed in the MQC. Its documentation[2] states that some sources with a strong influence from their host galaxies and QSO candidates are included along with confirmed core-dominated AGN. These objects exhibit, mostly, galaxy-like properties (with low radio emission levels, which might only be detected at redshift values close to $z = 0$), shifting their distributions to unusual shapes for AGN. In addition to that, and given that only around a 3% of the SDSS sources in our sample were detected in the radio bands when catalogued [49], and SDSS do not show this misshapen distribution, the distortion affects mostly radio-detected sources.

Even though this deviation from the expected distribution might affect, in some way, part of our results, we will keep the sources that produce it in our calculations. As mentioned in Section 1, part of the aims of this work is test whether an ML model can deliver reasonable results without discarding, or modifying, a large fraction of the intital dataset.

We also calculated colours for some of the bands. We computed g - r, r - i, i - z, z - y, g - i, W1 - W2, W2 - W3, W3 - W4, J - H, H - K, and FUV - NUV. In addition, following the results from Nakoneczny et al. [21], D'Isanto et al. [50], who studied different combinations of features and their positive impact on the prediction of redshifts, we have constructed ratios of magnitudes. The created quantities are r/z, i/y, W1/W3, W1/W4, W2/W4, J/K, and FUV/K. Finally, we included two indicators, in the form of a boolean flag, showing whether a source has a measurement on any radio band (LOFAR or TGSS) or on X-ray (Full band in XMM-Newton).

2.2. Methods

2.2.1. Data Preparation

Redshift values have a logarithmic behaviour when compared to the time passed—and distance travelled—between two values. A unit difference at low redshift has not the same significance as a unit difference at high redshift (i.e., early epochs). Given that, ultimately, redshifts can be used to determine distances, and times, from the observer to a given source, it is useful to make this quantity comparable to linear measurements. Thus, to overcome this non-linearity and, at the same time, establish a procedure to contrast predictions and real values, all comparisons will be normalised by the real redshift as follows:

$$\Delta z^N = \frac{z_{\text{Predicted}} - z_{\text{True}}}{1 + z_{\text{True}}} = \frac{\Delta z}{1 + z_{\text{True}}}. \tag{1}$$

Using these two quantities, Δz and Δz_N, it is possible to define a set of metrics to assess the quality of the prediction the developed models can achieve. First, we can define the standard deviation between the true, original redshift and the predicted value.

$$\sigma_{\Delta z} = \sqrt{\frac{1}{N} \sum_i^N \Delta z^2}. \tag{2}$$

In the same way, the value Δz^N can be used instead of Δz, giving rise to the normalised standard deviation, $\sigma_{\Delta z}^N$.

Alternatively, the redshift deviations can be used directly to create the median absolute deviation (MAD),

$$\sigma_{MAD} = 1.48 \times \text{median}|\Delta z|, \tag{3}$$

or the normalised MAD (NMAD) with the weighted redshift deviations, Δz^N.

Another quantity used to evaluate the predictions is the fraction of outliers, η. It represents the number of predictions that are too far away from the true value over the total number of prediction. There are several ways to define this value [51–54]. We will make use of the interpretation by Hildebrandt et al. [51], which considers all predictions that fulfil the following condition to be outliers:

$$\left| \Delta z^N \right| > 0.15. \tag{4}$$

Using both the standard and normalised differences between redshift values can allow us to analyse the results of our predictions from two points of view: from a purely statistical standpoint (using the standard difference), and a physically-motivated perspective, with the use of the normalised redshift difference. Both approaches can be useful to reach a better understanding of the behaviour of the used models.

For this work, we have analysed our data using the Regression module of the Python package `PyCaret`[3] [55]. It can create a full pipeline for the use of our dataset and has enough options to change its parameters as needed.

The first step of data preparation is imputation. A large fraction of ML models cannot be used with missing data. For this reason, several methods have been devised to impute missing values (for a review on data imputation, see, for instance, Ref. [56]). In our dataset, several features have a large fraction of empty spaces. A distribution of empty entries, prepared with the software `missingno` [57], can be seen in Figure 3. It is possible to see that radio and X-ray features have the largest number of empty values.

We imputed each magnitude with its detection limit and propagated those values for colours and ratios, assuming that empty entries are faint enough to be detected by each instrument. Thereafter, and within the `PyCaret` frame, we further removed features based on their influence over the prediction. We applied the Boruta method [58], discarding a feature if it behaves better than an aleatory version of itself. The final list of used features

is seen in Table 5. The remaining features are re-scaled to have a mean value of $\mu = 0$ and a standard deviation of $\sigma = 1$ and, afterwards, power-transformed to resemble a Gaussian distribution using the Yeo-Johnson method [59]. The use of re-scaling steps helped our models to improve their results over training with the original features. No further modifications were applied to the data. Thus, no corrections are applied for obscuration, AGN variability, host galaxy morphology, or other properties.

For the validation of our ML model, we have set aside a 10% of the full dataset. From the remaining 90%, 70% was used for training, and 30% for model testing. The same distribution of sources, which was created randomly, was used throughout the full study. Following the conventions used by `PyCaret`, the validation sub-set is the only fraction of the data which is not used for the training stage.
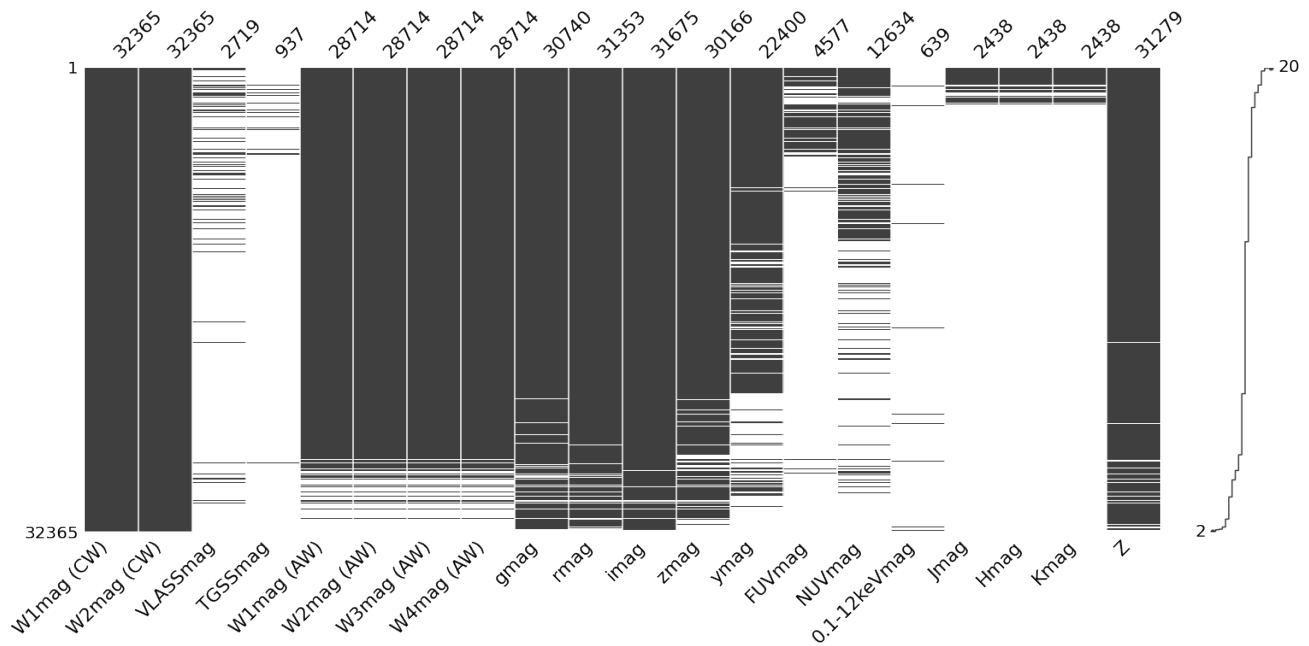

Figure 3. Distribution of empty values in HETDEX dataset. Each column shows the data from one feature and dark spaces indicate rows with a valid entry. The number of valid entries per feature is seen on top of each column. The dark line in the right side of the plot shows how many measurements each source in the dataset has. For clarity, sources have been sorted by number of entries, not affecting further results.

2.2.2. Model Selection and Stacking

With the help of `PyCaret`, we run simple realisations of a list of known ML model and selected, as meta-learner, the model with the best score (σ_z^N, see Section 2.2.1). After these tests, we stacked the four models with the following best metrics. Model stacking takes the results (predictions) from several models and adds them as new features for the meta-learner. In this way, the meta-learner can use the properties and advantages of the remaining models as a guidance for its own training and improve the prediction results. Furthermore, stacking can help improving the overall scores of the predictions. The stacked model was trained using 10-fold Cross Validation. The metrics of the training of the base and meta learners, along with those from the stacked model are presented in Table 2.

Table 2. Model Stacking results. Only σ_{NMAD} was used to rank the models and select base and meta learners (see Section 2.2.2). Stacked Train refers to the use of the stacked model in the training set and Stacked Train+Test to the same model in the union of training and test sets.

	Random Forest	Extra Trees	CatBoost	LightGBM	XGBoost	Stacked Train	Stacked Train + Test
σ_{NMAD}	0.1040	0.1079	0.1225	0.1251	0.1295	0.0971	0.1000
σ_z^N	0.4639	0.4608	0.4587	0.4656	0.4771	0.4495	0.4445

3. Results

3.1. Redshift Prediction

For the model stacking, we have chosen, as base models, `Extra Trees` (Extremely Randomised Trees, [60]), `CatBoost` [61,62], `LightGBM` [63], and `XGBoost` [64]. A `Random Forest` regression model [65] was used as meta-learner. From the 10-fold Cross Validation training, we have obtained a value of $\sigma_{NMAD} = 0.0971 \pm 0.0027$ (see Table 2, where the uncertainty value corresponds to the standard deviation of the Cross Validation instances). This is in the order of a one hundredth of a scaled redshift unit, improving upon the results from the individual models. In addition, when including the test set in the training of the model, the normalised standard deviation is $\sigma_{NMAD} = 0.1000$.

Figure 4a shows the prediction values for the validation set. The density of the plotted points, with higher values shown as a darker colours, shows that a large fraction of the predictions are close to the $y = x$ line. Additionally, the outlier fraction (Equation (4)) for the HETDEX Spring Field validation sample is $\eta = 21.87\%$. The results of the prediction over the test and validation sets are summarised in Table 3.
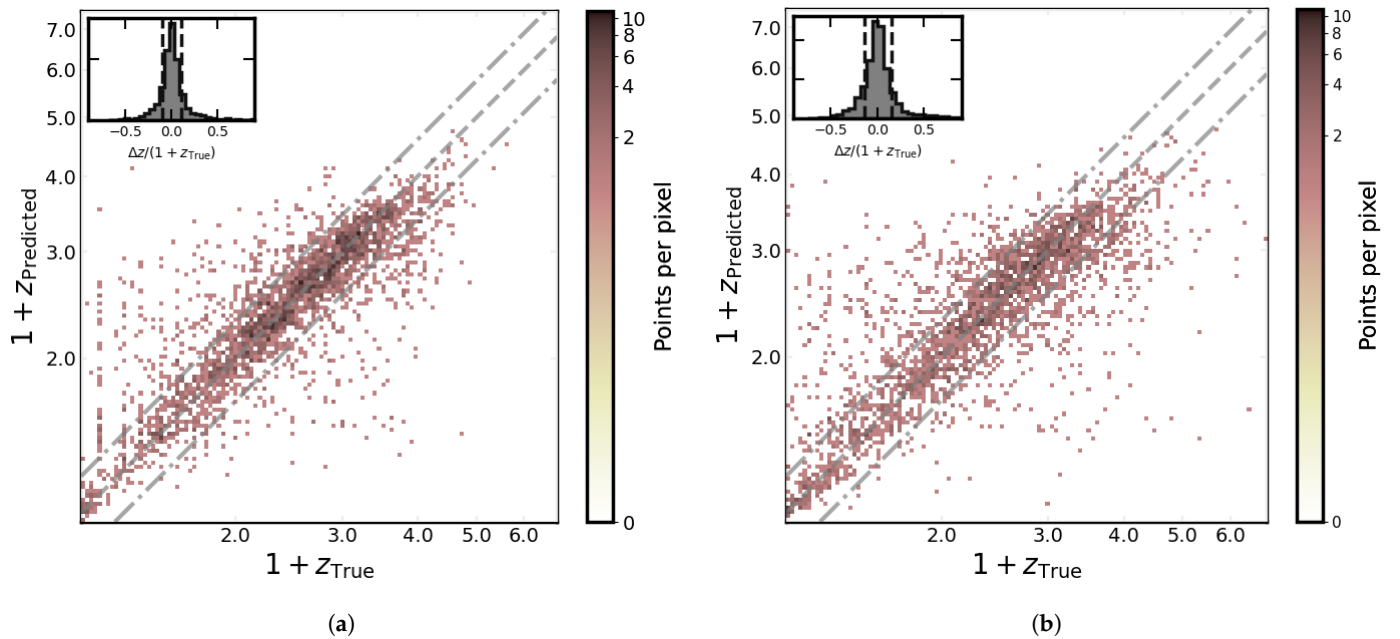

(a)

(b)

Figure 4. Distribution of predicted redshifts as a function of the original redshift values from the validation sample and Stripe 82 Field. Each square represents the number of sources as colour-coded in the colour bar. Diagonal, dashed line represents the $x = y$ relation and the dotted dashed lines show the zone of outliers. The panel in the upper-left side of each figure shows the distribution of Δz^N values from the prediction. (**a**) HETDEX Spring Field validation set; (**b**) Stripe 82 Field sample.

Table 3. Results from the application of the model to the Test and Validation sets, to the full Stripe 82 sample, and to the cross-match between our sample and the X-ray sources from Ananna et al. [66] (see Section 4.1).

	HETDEX Test Set	HETDEX Validation Set	Stripe 82 Test Set	Stripe 82 Ananna+17
σ_{MAD}	0.1392	0.2118	0.2854	0.2287
σ_{NMAD}	0.0594	0.0906	0.1197	0.1122
σ_z	0.2756	0.4287	0.5528	0.3630
σ_z^N	0.1162	0.1986	0.2501	0.1834
η	0.1158	0.2187	0.2972	0.2429

3.2. Prediction in Stripe 82 Field

To avoid possible biases derived from predicting on the same type of data as that used in training, and to test the prediction capabilities of our model, we applied it in data from a different area of the sky. In this case, we selected the SDSS Stripe 82 Field. We gathered the same data as described in Section 2.1. The main difference is that this field is not covered by the LoTSS-DR1 Survey. Thus, the selected area is defined by the coverage of the VLA SDSS Stripe 82 Survey [67]. This is to mimic the use, as with the HETDEX Spring Field, of an area covered by a radio survey. The VLA-Stripe 82 Survey covers an area over 92 deg^2 with a median rms noise of 52 μJ/beam and an angular resolution of 1.8″. We have selected this field because of the high-quality measurements it hosts, and thorough studies on AGN over its area. The sample we have produced has 369,093 detected sources and 2941 of them have been labelled as AGN by the MQC. Additionally, 111 sources have been defined as QSO candidates.

In Figure 4b, the results of the redshift predictions, along with the original values, for Stripe 82 are presented. Results from Stripe 82, shown in Table 3, resemble those of the HETDEX Spring Field, hinting the possibility of, as long as the needed wavebands are available, using the trained models in areas of the sky which are not related to the training sample.

In addition, and even though all metric results are better in the initial HETDEX Spring sample, differences with Stripe 82 are on the range of 7–8%. These deviations are small enough to be caused by statistical variations among both fields. In the case of the outlier fraction, it is around 30%, and 8 percentage points higher than with the primary sample.

4. Discussion

4.1. Previous Results

As a way to assess our results, it is possible to compare them to previous redshift determinations. This is the case of Ananna et al. [66]. They used multiwavelength data from 5961 X-ray-detected AGN in the Stripe 82 Field with $z \leq 3.0$ and, from fitting SED models, they computed photometric redshifts. From their Table 7, a value of $\sigma_{NMAD} = 0.0602$ is quoted for their full sample, which is in line with our prediction in the Stripe 82 Field ($\sigma_{NMAD} = 0.1197$). In addition, an outlier fraction of 13.69% is achieved, less than half of what is obtained using our stacked model in the same area. It is possible to select, from our sample, the sources with a counterpart in the Ananna et al. [66] sample and apply our model to them. Using a matching radius of 2″, 221 sources are selected, reaching values of $\sigma_{NMAD} = 0.1122$ and $\eta = 0.2429$. If we do the same exercise, selecting the results from the SED-fitting redshift determination, their values are $\sigma_{NMAD} = 0.0648$ and $\eta = 0.2048$. Full results for this sub-sample are shown in Table 4.

To contrast our results with previous ML implementations, we can take the work from Curran et al. [68], who compared the results of applying deep learning, decision trees, and k-nearest neighbours regression to predict redshift values for 100,000 SDSS DR12 QSO with accurate spectroscopic redshifts. Results are presented in Table 4.

Table 4. Results from previous works. First column: full X-ray selected sample quoted from Ananna et al. [66]. Second column: selection of sources from Ananna et al. [66] that have a match in our sample. Following columns: result of application of k-Nearest Neigbours (KN), Decision Tree Regression (DT), and Deep Learning (DL) models as quoted from Curran et al. [68].

	Stripe 82 Full Ananna+17	Stripe 82 Match Ananna+17	SDSS KN Curran+2021	SDSS DT Curran+2021	SDSS DL Curran+2021
σ_{MAD}	$\cdots$	0.1336	0.2360	0.1290	0.0920
σ_{NMAD}	0.0602	0.0648	0.0500	0.0580	0.0420
σ_z	$\cdots$	0.5435	0.2360	0.3330	0.2350
σ_z^N	$\cdots$	0.2766	0.1210	0.1600	0.1100
η	0.1369	0.2048	$\cdots$	$\cdots$	$\cdots$

When comparing our results (Table 3) with the outputs from Curran et al. [68], we note that the metrics for our Validation set are 20–40% higher and those from the Stripe 82 Field, 40–60% higher than theirs. This is a consequence of our decision of not cleaning our training set, mimicking the conditions a large dataset might present. They, in contrast, have trained their models with sources that have full coverage on the bands they selected, avoiding the use of imputation. Moreover, since they have used large SDSS sample, the properties of QSO among them are more homogeneous than that of the present work, leading to improved prediction results.

Comparison with previous works, using traditional template-based and ML photometric redshift determination methods, highlights the prospective scenarios to apply our model. Rather than selecting a very small area with the right conditions, we can use the model here presented on large regions with incomplete coverage, rising the likelihood of obtaining objects with specific resdhift values.

4.2. Feature Importances

Feature importances from our model are listed in Table 5. The values are provided by the model itself, and they have been calculated as the mean decrease on impurity for the ensemble of trees. We can see that the features with the highest importances are those coming from the CatWISE catalogue. After them, quantities are derived from Pan-STARRS. In addition, finally, those obtained from AllWISE, and GALEX observations suggest a very low impact in the model training and the predictions derived from it.

Table 5. Features used by our redshift prediction model and their importances.

Feature	Importance	Feature	Importance	Feature	Importance
W1 - W2 (CW)	87.381	z - y	37.084	FUV - NUV	11.338
W1 (CW)	82.759	W1/W3 (AW)	33.207	FUV/K	8.886
g - i	70.617	i/y	33.081	FUV	7.202
g	55.787	W2/W4 (AW)	29.196	K	5.484
W2 - W3 (AW)	53.919	i - z	28.647	J - H	2.817
r/z	52.251	W4 (AW)	26.392	J/K	2.803
y	49.234	W3 - W4 (AW)	24.898	H - K	2.771
r - i	46.451	NUV	23.296		

Entries from CatWISE have the largest amount of relevant, non-repetitive information from all features. Despite the different nature of the used features, i.e., magnitudes, colours, ratios, there is no clear preference of one kind over the others. The main factor to have high importance is the fraction of sources with a measurement in the studied feature. This distribution also reinforces the results from Ref. [50], who established that is possible to use combinations of magnitudes other than colours and train, successfully, ML models.

Table 5 also gives information on the features that can be discarded from the model training without having a high influence on the predicted values (features with data from

2MASS and GALEX). Finally, it is important to stress that, in this work, we have not discarded data based upon the feature importances.

4.3. Shapley Explanations

Shapley values were obtained using the Tree-based module of the Python package SHAP[4] [69,70]. In Figure 5, features are sorted by decreasing median Shapley values.

The quantity with the highest Shapley value is related to the base observations. However, from the distribution of values in the horizontal axis for the W1 magnitude, it is possible to see that its large dispersion implies that its influence on predictions can drive the final redshift either to low or high values. This is in contrast with, for instance, the g - i colour. Its Shapley values might be close to zero, indicating that it does not have impact on the redshfit prediction. The values can be higher than zero, as well, driving the predictions to high redshift values.
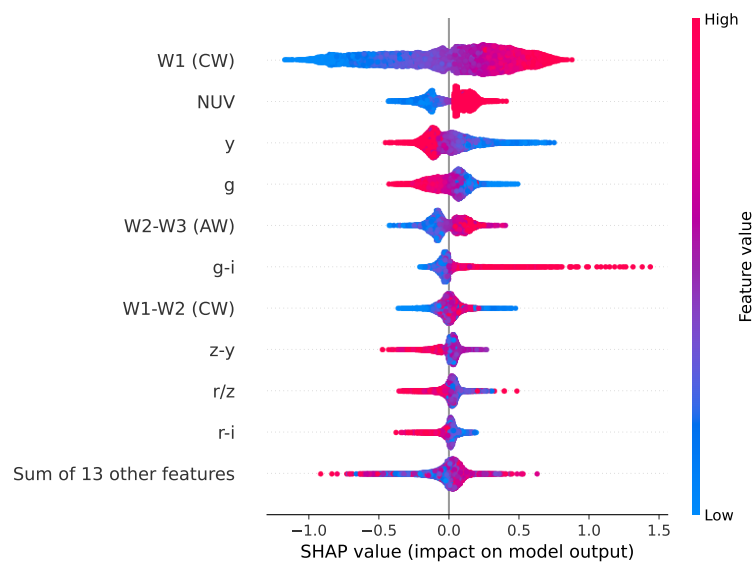
Figure 5. Shapley values for the ten features with the highest median Shapley numbers in our redshift prediction model. Each row corresponds to one feature. Colour map indicates the value of the feature for each source. Features are sorted by median Shapley value. Last row shows the sum of the 13 remaining features used for the prediction. Feature values of points to the left of vertical, grey line have a positive impact on the model output, i.e., redshift will tend to be higher. Points close to the vertical line show a limited impact on the redshift prediction.

Most of the remaining features show Shapley values clustered around 0.0, and a small sub-sample deviates from this and has a noteworthy influence on predictions.

The feature with the second highest median Shapley values is the NUV magnitude from GALEX. From Figure 3, it is possible to see that this feature exhibits a very high fraction of empty entries. That implies that most of sources have an imputed NUV magnitude. This distribution is present in Figure 5. Therefore, all imputed magnitudes make the redshift prediction go up, and all measured magnitudes make it go down. Although this behaviour might seem anomalous, it has its roots on the fact that very few high-redshift sources are detectable by GALEX.

Being able to retrieve these interpretations is one of the advantages of using Shapley values from a prediction model. It is possible to understand whether certain range of values of a feature can make a prediction go up or down. This differs from feature importances, which allow an average view of the impact of a feature over the complete trained model.

Despite their differences, feature importances and Shapley values can help understand the impact that measurements in different wavelengths can have over the understanding and prediction of redshift values of AGN. In particular, and given the relevance and

high-quality observations that future radio surveys and observatories will deliver, adding direct measurements (e.g., Ref. [71]) or features derived from them might be highly beneficial when focusing the search on high-redshift objects. The latter might be the case with already-known quantities, such as radio loudness or radio spectral indices. These properties can provide indications on the radio emission [72] and its relation with other wavelengths [73,74].

5. Conclusions

In this work, we trained several Machine Learning models to predict, from a sample of infrared-detected AGN—and their multiwavelength counterparts—their redshift value.

Sources were obtained from CatWISE2020 catalogue and counterpart measurements were obtained from AllWISE, Pan-STARRS, LOFAR, GMRT, VLASS, GALEX, 2MASS, and XMM-NEWTON observations and surveys. All of the sources are located at the HETDEX Spring Field.

Using of the `PyCaret` Python package as a framework, we stacked four different models with a meta-learner. The application of model to the validation set lead a median redshift error on the prediction of $\sigma_z^N = 0.1986$ and an outlier fraction of $\eta = 21.87\%$. This goes in line with previous results, taking into account that no major cleaning procedure was performed into the dataset.

To further test the power of our model, we applied it to a separate catalogue of AGN located in the Stripe 82 Field, and the median redshift error was $\sigma_z^N = 0.2501$ and an outlier fraction of $\eta = 29.72\%$.

To understand the influence of the different features included in the model, Shapley values were calculated for the training sub-set. The features from WISE and from Pan-STARRS show the highest median Shapley values, mirroring the fact that these features have the lowest number of imputed entries.

The results presented in this work stress the benefits of using ML as an initial approach to derive redshift predictions for AGN. Using a fraction of the time a template-based photometric redshift determination tool might take, ML can give redshift predictions with a high confidence level which can lead to further studies of selected sources. This advantage might become critical to the use of current and future large-area surveys—with radio surveys being a major example, which need to extract information from several millions of sources within an appropriate amount of time.

Even though some of the results obtained in this work do not show a considerable improvement from previous studies, it is relevant to emphasise that our work was aimed to extract predictions using datasets without large amounts of preparation, i.e., feature engineering. This implies that it is possible to use a very heterogeneous group of datasets (with different sensitivities, resolutions, etc.) and obtain useful predictions from them without the need of cleaning and reducing the number of used sources in each catalogue.

Our model can be further improved using future surveys which will cover large areas with very deep observations. One such survey is Data Release 2 of the LoTSS survey, which will be released in the near future. It will cover 5720 deg^2 in the northern sky with similar sensitivities as DR1 [75]. If assuming the same AGN density as in LoTSS DR1 (see Section 2.1, with 32,365 AGN in 424 deg^2), DR2 is expected to deliver 436,622 AGN from its area. This will allow us training a redshift prediction model with a number of sources one order of magnitude larger, improving its accuracy dramatically by capturing the properties from a larger parameter space. This improvement can be also analysed in terms of cosmic variance. Following the results by Ref. [76], DR1 from the LoTSS survey will be subject to a cosmic variance between 10 and 20%. In addition, extrapolating the curve from their Figure 6, DR2 will make this value go below 10%. Only from this improvement, we might expect to achieve a better training for a prediction model. AGN might present variability on their observations with different timescales [77,78], which might impact the observed properties of the used datasets. These variations can increase the fraction of outliers in different ranges [79].

Additional sources of improvement in the results are related to the treatment of the missing values in our catalogue. Devising more advanced imputation methods, which can take into account the distribution of measured values in one feature and their relation to the rest of features, might refine our results. Related to this, some features have a low fraction of measured values, adding little information to the models. Discarding these features also might reduce the fraction of outliers. Apart from the data treatment, further improvements might be achieved if the intrinsic time variability of AGN is taken into account.

The used model might arrive to better results creating several instances of data subsets. Using different combinations of sources for training, test, and validation might have an impact on how the model arrives to separate predictions.

With all these advantages, the model described in this article can be used as part of a full pipeline which might be able to predict the presence of AGN in a large-area field. In addition, for the predicted AGN, we predict their redshift values, among other properties, e.g., radio detectability. This might allow the creation of catalogues with high-redshift Radio Galaxies from datasets covering large areas.

of data obtained from the 4XMM XMM-Newton serendipitous source catalogue compiled by the 10 institutes of the XMM-Newton Survey Science Centre selected by ESA. This publication makes use of data products from the Two Micron All Sky Survey, which is a joint project of the University of Massachusetts and the Infrared Processing and Analysis Center/California Institute of Technology, funded by the National Aeronautics and Space Administration and the National Science Foundation. This research has made use of the VizieR catalogue access tool, CDS, Strasbourg, France (DOI: 10.26093/cds/vizier). The original description of the VizieR service was published in Ref. [81]. This research made use of Astropy[6] , a community-developed core Python package for Astronomy [82,83] and TOPCAT[7] [84]. This research has made use of NASA's Astrophysics Data System.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AGN	Active Galactic Nuclei
QSO	Quasi Stellar Object
ML	Machine Learning
RG	Radio Galaxy
EoR	Epoch of Reionisation
CW	CatWISE2020 Catalogue
AW	AllWISE Catalogue

Notes

1   http://quasars.org/milliquas.htm (accessed on 3 May 2021).
2   http://quasars.org/Milliquas-ReadMe.txt (accessed on 25 October 2021).
3   https://pycaret.org (accessed on 23 October 2021).
4   https://github.com/slundberg/shap (accessed on 18 October 2021).
5   https://lofar-surveys.org/ (accessed on 3 August 2021).
6   https://www.astropy.org (accessed on 23 July 2021).
7   http://www.star.bris.ac.uk/~mbt/topcat/ (accessed on 29 July 2021).

References

1.   Padovani, P.; Alexander, D.M.; Assef, R.J.; De Marco, B.; Giommi, P.; Hickox, R.C.; Richards, G.T.; Smolčić, V.; Hatziminaoglou, E.; Mainieri, V.; et al. Active galactic nuclei: What's in a name? *Astron. Astrophys. Rev.* **2017**, *25*, 2. [CrossRef]
2.   Heckman, T.M.; Best, P.N. The Coevolution of Galaxies and Supermassive Black Holes: Insights from Surveys of the Contemporary Universe. *Annu. Rev. Astron. Astrophys.* **2014**, *52*, 589–660. [CrossRef]
3.   McGreer, I.D.; Becker, R.H.; Helfand, D.J.; White, R.L. Discovery of a z = 6.1 Radio-Loud Quasar in the NOAO Deep Wide Field Survey. *Astrophys. J.* **2006**, *652*, 157–162. [CrossRef]
4.   Kuźmicz, A.; Jamrozy, M. Giant Radio Quasars: Sample and Basic Properties. *Astrophys. J.* **2021**, *253*, 25. [CrossRef]
5.   Delhaize, J.; Heywood, I.; Prescott, M.; Jarvis, M.J.; Delvecchio, I.; Whittam, I.H.; White, S.V.; Hardcastle, M.J.; Hale, C.L.; Afonso, J.; et al. MIGHTEE: Are giant radio galaxies more common than we thought? *Mon. Not. R. Astron. Soc.* **2021**, *501*, 3833–3845. [CrossRef]
6.   Lal, D.V. The Discovery of a Remnant Radio Galaxy in A2065 Using GMRT. *Astrophys. J.* **2021**, *915*, 126. [CrossRef]
7.   Amarantidis, S.; Afonso, J.; Messias, H.; Henriques, B.; Griffin, A.; Lacey, C.; Lagos, C.d.P.; Gonzalez-Perez, V.; Dubois, Y.; Volonteri, M.; et al. The first supermassive black holes: Indications from models for future observations. *Mon. Not. R. Astron. Soc.* **2019**, *485*, 2694–2709. [CrossRef]
8.   Thomas, N.; Davé, R.; Jarvis, M.J.; Anglés-Alcázar, D. The radio galaxy population in the SIMBA simulations. *Mon. Not. R. Astron. Soc.* **2021**, *503*, 3492–3509. [CrossRef]
9.   Bonaldi, A.; Bonato, M.; Galluzzi, V.; Harrison, I.; Massardi, M.; Kay, S.; De Zotti, G.; Brown, M.L. The Tiered Radio Extragalactic Continuum Simulation (T-RECS). *Mon. Not. R. Astron. Soc.* **2019**, *482*, 2–19. [CrossRef]
10.   Prandoni, I.; Seymour, N. Revealing the Physics and Evolution of Galaxies and Galaxy Clusters with SKA Continuum Surveys. In Proceedings of the Advancing Astrophysics with the Square Kilometre Array (AASKA14), Giardini Naxos, Italy, 9–13 June 2014; p. 67.
11.   Inayoshi, K.; Visbal, E.; Haiman, Z. The Assembly of the First Massive Black Holes. *Annu. Rev. Astron. Astrophys.* **2020**, *58*, 27–97. [CrossRef]

12. Ross, N.P.; Cross, N.J.G. The near and mid-infrared photometric properties of known redshift z ≥ 5 quasars. *Mon. Not. R. Astron. Soc.* **2020**, *494*, 789–803. [CrossRef]
13. Miley, G.; De Breuck, C. Distant radio galaxies and their environments. *Astron. Astrophys. Rev.* **2008**, *15*, 67–144. [CrossRef]
14. Helfand, D.J.; White, R.L.; Becker, R.H. The Last of FIRST: The Final Catalog and Source Identifications. *Astrophys. J.* **2015**, *801*, 26. [CrossRef]
15. Norris, R.P.; Hopkins, A.M.; Afonso, J.; Brown, S.; Condon, J.J.; Dunne, L.; Feain, I.; Hollow, R.; Jarvis, M.; Johnston-Hollitt, M.; et al. EMU: Evolutionary Map of the Universe. *Publ. Astron. Soc. Aust.* **2011**, *28*, 215–248. [CrossRef]
16. Gordon, Y.A.; Boyce, M.M.; O'Dea, C.P.; Rudnick, L.; Andernach, H.; Vantyghem, A.N.; Baum, S.A.; Bui, J.P.; Dionyssiou, M. A Catalog of Very Large Array Sky Survey Epoch 1 Quick Look Components, Sources, and Host Identifications. *Res. Notes Am. Astron. Soc.* **2020**, *4*, 175. [CrossRef]
17. Shimwell, T.W.; Tasse, C.; Hardcastle, M.J.; Mechev, A.P.; Williams, W.L.; Best, P.N.; Röttgering, H.J.A.; Callingham, J.R.; Dijkema, T.J.; de Gasperin, F.; et al. The LOFAR Two-metre Sky Survey. II. First data release. *Astron. Astrophys.* **2019**, *622*, A1. [CrossRef]
18. Singh, V.; Beelen, A.; Wadadekar, Y.; Sirothia, S.; Ishwara-Chandra, C.H.; Basu, A.; Omont, A.; McAlpine, K.; Ivison, R.J.; Oliver, S.; et al. Multiwavelength characterization of faint ultra steep spectrum radio sources: A search for high-redshift radio galaxies. *Astron. Astrophys.* **2014**, *569*, A52. [CrossRef]
19. Williams, W.L.; Calistro Rivera, G.; Best, P.N.; Hardcastle, M.J.; Röttgering, H.J.A.; Duncan, K.J.; de Gasperin, F.; Jarvis, M.J.; Miley, G.K.; Mahony, E.K.; et al. LOFAR-Boötes: Properties of high- and low-excitation radio galaxies at 0.5 < z < 2.0. *Mon. Not. R. Astron. Soc.* **2018**, *475*, 3429–3452. [CrossRef]
20. Capetti, A.; Brienza, M.; Baldi, R.D.; Giovannini, G.; Morganti, R.; Hardcastle, M.J.; Rottgering, H.J.A.; Brunetti, G.F.; Best, P.N.; Miley, G. The LOFAR view of FR 0 radio galaxies. *Astron. Astrophys.* **2020**, *642*, A107. [CrossRef]
21. Nakoneczny, S.J.; Bilicki, M.; Pollo, A.; Asgari, M.; Dvornik, A.; Erben, T.; Giblin, B.; Heymans, C.; Hildebrandt, H.; Kannawadi, A.; et al. Photometric selection and redshifts for quasars in the Kilo-Degree Survey Data Release 4. *Astron. Astrophys.* **2021**, *649*, A81. [CrossRef]
22. Wenzl, L.; Schindler, J.T.; Fan, X.; Andika, I.T.; Bañados, E.; Decarli, R.; Jahnke, K.; Mazzucchelli, C.; Onoue, M.; Venemans, B.P.; et al. Random Forests as a Viable Method to Select and Discover High-redshift Quasars. *Astron. J.* **2021**, *162*, 72. [CrossRef]
23. Ma, Z.; Xu, H.; Zhu, J.; Hu, D.; Li, W.; Shan, C.; Zhu, Z.; Gu, L.; Li, J.; Liu, C.; et al. A Machine Learning Based Morphological Classification of 14,245 Radio AGNs Selected from the Best-Heckman Sample. *Astrophys. J.* **2019**, *240*, 34. [CrossRef]
24. Lukic, V.; Brüggen, M.; Mingo, B.; Croston, J.H.; Kasieczka, G.; Best, P.N. Morphological classification of radio galaxies: Capsule networks versus convolutional neural networks. *Mon. Not. R. Astron. Soc.* **2019**, *487*, 1729–1744. [CrossRef]
25. Mostert, R.I.J.; Duncan, K.J.; Röttgering, H.J.A.; Polsterer, K.L.; Best, P.N.; Brienza, M.; Brüggen, M.; Hardcastle, M.J.; Jurlin, N.; Mingo, B.; et al. Unveiling the rarest morphologies of the LOFAR Two-metre Sky Survey radio source population with self-organised maps. *Astron. Astrophys.* **2021**, *645*, A89. [CrossRef]
26. Vardoulaki, E.; Jiménez Andrade, E.F.; Delvecchio, I.; Smolčić, V.; Schinnerer, E.; Sargent, M.T.; Gozaliasl, G.; Finoguenov, A.; Bondi, M.; Zamorani, G.; et al. FR-type radio sources at 3 GHz VLA-COSMOS: Relation to physical properties and large-scale environment. *Astron. Astrophys.* **2021**, *648*, A102. [CrossRef]
27. Burhanudin, U.F.; Maund, J.R.; Killestein, T.; Ackley, K.; Dyer, M.J.; Lyman, J.; Ulaczyk, K.; Cutter, R.; Mong, Y.L.; Steeghs, D.; et al. Light-curve classification with recurrent neural networks for GOTO: Dealing with imbalanced data. *Mon. Not. R. Astron. Soc.* **2021**, *505*, 4345–4361. [CrossRef]
28. Saz Parkinson, P.M.; Xu, H.; Yu, P.L.H.; Salvetti, D.; Marelli, M.; Falcone, A.D. Classification and Ranking of Fermi LAT Gamma-ray Sources from the 3FGL Catalog using Machine Learning Techniques. *Astrophys. J.* **2016**, *820*, 8. [CrossRef]
29. Chiaro, G.; Salvetti, D.; La Mura, G.; Giroletti, M.; Thompson, D.J.; Bastieri, D. Blazar flaring patterns (B-FlaP) classifying blazar candidate of uncertain type in the third Fermi-LAT catalogue by artificial neural networks. *Mon. Not. R. Astron. Soc.* **2016**, *462*, 3180–3195. [CrossRef]
30. Xiao, H.B.; Cao, H.T.; Fan, J.H.; Costantin, D.; Luo, G.Y.; Pei, Z.Y. Efficient Fermi source identification with machine learning methods. *Astron. Comput.* **2020**, *32*, 100387. [CrossRef]
31. Wang, C.; Bai, Y.; López-Sanjuan, C.; Yuan, H.; Wang, S.; Liu, J.; Sobral, D.; Baqui, P.O.; Martín, E.L.; Galarza, C.A.; et al. J-PLUS: Support Vector Machine Applied to STAR-GALAXY-QSOClassification. *arXiv* **2021**, arXiv:2106.12787.
32. Li, Y.; Ni, Y.; Croft, R.A.C.; Di Matteo, T.; Bird, S.; Feng, Y. AI-assisted superresolution cosmological simulations. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2022038118. [CrossRef] [PubMed]
33. Ball, N.M.; Brunner, R.J. Data Mining and Machine Learning in Astronomy. *Int. J. Mod. Phys. D* **2010**, *19*, 1049–1106. [CrossRef]
34. Baron, D. Machine Learning in Astronomy: A practical overview. *arXiv* **2019**, arXiv:1904.07248.
35. Goebel, R.; Chander, A.; Holzinger, K.; Lecue, F.; Akata, Z.; Stumpf, S.; Kieseberg, P.; Holzinger, A. Explainable ai: The new 42? In Proceedings of the International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Hamburg, Germany, 27–30 August 2018; Springer International Publishing: Cham, Switzerland, 2018; pp. 295–303.
36. Roscher, R.; Bohn, B.; Duarte, M.F.; Garcke, J. Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Access* **2020**, *8*, 42200–42216. [CrossRef]
37. Shapley, L.S. A Value for n-Person Games. In *Contributions to the Theory of Games (AM-28), Volume II*; Princeton University Press: Princeton, NJ, USA, 1953; Volume 1, pp. 307–318. [CrossRef]

38. Molnar, C. Interpretable Machine Learning. 2019. Available online: https://christophm.github.io/interpretable-ml-book/ (accessed on 4 May 2021).
39. Marocco, F.; Eisenhardt, P.R.M.; Fowler, J.W.; Kirkpatrick, J.D.; Meisner, A.M.; Schlafly, E.F.; Stanford, S.A.; Garcia, N.; Caselden, D.; Cushing, M.C.; et al. The CatWISE2020 Catalog. *Astrophys. J.* **2021**, *253*, 8. [CrossRef]
40. Fernique, P.; Boch, T.; Donaldson, T.; Durand, D.; O'Mullane, W.; Reinecke, M.; Taylor, M. MOC—HEALPix Multi-Order Coverage map Version 1.0. *arXiv* **2015**, arXiv:1505.02937.
41. Flewelling, H.A.; Magnier, E.A.; Chambers, K.C.; Heasley, J.N.; Holmberg, C.; Huber, M.E.; Sweeney, W.; Waters, C.Z.; Calamida, A.; Casertano, S.; et al. The Pan-STARRS1 Database and Data Products. *Astrophys. J.* **2020**, *251*, 7. [CrossRef]
42. Bianchi, L.; Shiao, B.; Thilker, D. Revised Catalog of GALEX Ultraviolet Sources. I. The All-Sky Survey: GUVcat_AIS. *Astrophys. J.* **2017**, *230*, 24. [CrossRef]
43. Intema, H.T.; Jagannathan, P.; Mooley, K.P.; Frail, D.A. The GMRT 150 MHz all-sky radio survey. First alternative data release TGSS ADR1. *Astron. Astrophys.* **2017**, *598*, A78. [CrossRef]
44. Traulsen, I.; Schwope, A.D.; Lamer, G.; Ballet, J.; Carrera, F.J.; Ceballos, M.T.; Coriat, M.; Freyberg, M.J.; Koliopanos, F.; Kurpas, J.; et al. The XMM-Newton serendipitous survey. X. The second source catalogue from overlapping XMM-Newton observations and its long-term variable content. *Astron. Astrophys.* **2020**, *641*, A137. [CrossRef]
45. Cutri, R.M.; Skrutskie, M.F.; van Dyk, S.; Beichman, C.A.; Carpenter, J.M.; Chester, T.; Cambresy, L.; Evans, T.; Fowler, J.; Gizis, J.; et al. 2MASS All Sky Catalog of Point Sources. 2003. Available online: https://ui.adsabs.harvard.edu/abs/2003tmc..book.....C/abstract (accessed on 29 May 2021).
46. Skrutskie, M.F.; Cutri, R.M.; Stiening, R.; Weinberg, M.D.; Schneider, S.; Carpenter, J.M.; Beichman, C.; Capps, R.; Chester, T.; Elias, J.; et al. The Two Micron All Sky Survey (2MASS). *Astron. J.* **2006**, *131*, 1163–1183. [CrossRef]
47. Cutri, R.M.; Wright, E.L.; Conrow, T.; Fowler, J.W.; Eisenhardt, P.R.M.; Grillmair, C.; Kirkpatrick, J.D.; Masci, F.; McCallon, H.L.; Wheelock, S.L.; et al. Explanatory Supplement to the AllWISE Data Release Products. 2013. Available online: https://ui.adsabs.harvard.edu/abs/2013wise.rept....1C (accessed on 29 May 2021).
48. Flesch, E.W. The Million Quasars (Milliquas) v7.2 Catalogue, now with VLASS associations. The inclusion of SDSS-DR16Q quasars is detailed. *arXiv* **2021**, arXiv:2105.12985.
49. Lyke, B.W.; Higley, A.N.; McLane, J.N.; Schurhammer, D.P.; Myers, A.D.; Ross, A.J.; Dawson, K.; Chabanier, S.; Martini, P.; Busca, N.G.; et al. The Sloan Digital Sky Survey Quasar Catalog: Sixteenth Data Release. *Astrophys. J.* **2020**, *250*, 8. [CrossRef]
50. D'Isanto, A.; Cavuoti, S.; Gieseke, F.; Polsterer, K.L. Return of the features. Efficient feature selection and interpretation for photometric redshifts. *Astron. Astrophys.* **2018**, *616*, A97. [CrossRef]
51. Hildebrandt, H.; Arnouts, S.; Capak, P.; Moustakas, L.A.; Wolf, C.; Abdalla, F.B.; Assef, R.J.; Banerji, M.; Benítez, N.; Brammer, G.B.; et al. PHAT: PHoto-z Accuracy Testing. *Astron. Astrophys.* **2010**, *523*, A31. [CrossRef]
52. Bernstein, G.; Huterer, D. Catastrophic photometric redshift errors: Weak-lensing survey requirements. *Mon. Not. R. Astron. Soc.* **2010**, *401*, 1399–1408. [CrossRef]
53. Henghes, B.; Pettitt, C.; Thiyagalingam, J.; Hey, T.; Lahav, O. Benchmarking and scalability of machine-learning methods for photometric redshift estimation. *Mon. Not. R. Astron. Soc.* **2021**, *505*, 4847–4856. [CrossRef]
54. Brescia, M.; Cavuoti, S.; Razim, O.; Amaro, V.; Riccio, G.; Longo, G. Photometric Redshifts With Machine Learning, Lights and Shadows on a Complex Data Science Use Case. *Front. Astron. Space Sci.* **2021**, *8*, 70. [CrossRef]
55. Ali, M. PyCaret: An Open Source, Low-Code Machine Learning Library in Python. PyCaret Version 2.3. 2020. Available online: https://www.pycaret.org (accessed on 23 October 2021).
56. Chattopadhyay, A.K. Incomplete Data in Astrostatistics. In *Wiley StatsRef: Statistics Reference Online*; American Cancer Society: Atlanta, GA, USA, 2017; pp. 1–12. [CrossRef]
57. Bilogur, A.; Samuelbr; Beutner, V.; Fandango, A.; Everson, B.; Chacreton, D.; Abahurire, E.J.; Mavroforakis, H.; Cruz, J.S.; Mahlke, M.; et al. ResidentMario/missingno: 0.5.0 maintenance release. *Zenodo* **2021**. [CrossRef]
58. Kursa, M.B.; Rudnicki, W.R. Feature Selection with the Boruta Package. *J. Stat. Softw. Artic.* **2010**, *36*, 1–13. [CrossRef]
59. Yeo, I.; Johnson, R.A. A new family of power transformations to improve normality or symmetry. *Biometrika* **2000**, *87*, 954–959. [CrossRef]
60. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42. [CrossRef]
61. Dorogush, A.V.; Gulin, A.; Gusev, G.; Kazeev, N.; Prokhorenkova, L.O.; Vorobev, A. Fighting biases with dynamic boosting. *arXiv* **2017**, *arXiv:1706.09516*.
62. Dorogush, A.V.; Ershov, V.; Gulin, A. CatBoost: Gradient boosting with categorical features support. *arXiv* **2018**, *arXiv:1810.11363*.
63. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
64. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; ACM: New York, NY, USA, 2016; pp. 785–794. [CrossRef]
65. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

66. Ananna, T.T.; Salvato, M.; LaMassa, S.; Urry, C.M.; Cappelluti, N.; Cardamone, C.; Civano, F.; Farrah, D.; Gilfanov, M.; Glikman, E.; et al. AGN Populations in Large-volume X-Ray Surveys: Photometric Redshifts and Population Types Found in the Stripe 82X Survey. *Astrophys. J.* **2017**, *850*, 66. [CrossRef]
67. Hodge, J.A.; Becker, R.H.; White, R.L.; Richards, G.T.; Zeimann, G.R. High-resolution Very Large Array Imaging of Sloan Digital Sky Survey Stripe 82 at 1.4 GHz. *Astron. J.* **2011**, *142*, 3. [CrossRef]
68. Curran, S.J.; Moss, J.P.; Perrott, Y.C. QSO photometric redshifts using machine learning and neural networks. *Mon. Not. R. Astron. Soc.* **2021**, *503*, 2639–2650. [CrossRef]
69. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 4765–4774.
70. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 2522–5839. [CrossRef]
71. Turner, R.J.; Drouart, G.; Seymour, N.; Shabala, S.S. RAiSERed: Radio continuum redshifts for lobed active galactic nuclei. *Mon. Not. R. Astron. Soc.* **2020**, *499*, 3660–3672. [CrossRef]
72. Zajaček, M.; Busch, G.; Valencia-S., M.; Eckart, A.; Britzen, S.; Fuhrmann, L.; Schneeloch, J.; Fazeli, N.; Harrington, K.C.; Zensus, J.A. Radio spectral index distribution of SDSS-FIRST sources across optical diagnostic diagrams. *Astron. Astrophys.* **2019**, *630*, A83. [CrossRef]
73. Laor, A.; Behar, E. On the origin of radio emission in radio-quiet quasars. *Mon. Not. R. Astron. Soc.* **2008**, *390*, 847–862. [CrossRef]
74. Laor, A.; Baldi, R.D.; Behar, E. What drives the radio slopes in radio-quiet quasars? *Mon. Not. R. Astron. Soc.* **2019**, *482*, 5513–5523. [CrossRef]
75. McKean, J.P.; Luichies, R.; Drabent, A.; Gürkan, G.; Hartley, P.; Lafontaine, A.; Prandoni, I.; Röttgering, H.J.A.; Shimwell, T.W.; Stacey, H.R.; et al. Gravitational lensing in LoTSS DR2: Extremely faint 144-MHz radio emission from two highly magnified quasars. *Mon. Not. R. Astron. Soc.* **2021**, *505*, L36–L40. [CrossRef]
76. Driver, S.P.; Robotham, A.S.G. Quantifying cosmic variance. *Mon. Not. R. Astron. Soc.* **2010**, *407*, 2131–2140. [CrossRef]
77. Wolf, C.; Meisenheimer, K.; Kleinheinrich, M.; Borch, A.; Dye, S.; Gray, M.; Wisotzki, L.; Bell, E.F.; Rix, H.W.; Cimatti, A.; et al. A catalogue of the Chandra Deep Field South with multi-colour classification and photometric redshifts from COMBO-17. *Astron. Astrophys.* **2004**, *421*, 913–936. [CrossRef]
78. Salvato, M.; Hasinger, G.; Ilbert, O.; Zamorani, G.; Brusa, M.; Scoville, N.Z.; Rau, A.; Capak, P.; Arnouts, S.; Aussel, H.; et al. Photometric Redshift and Classification for the XMM-COSMOS Sources. *Astrophys. J.* **2009**, *690*, 1250–1263. [CrossRef]
79. Matute, I.; Márquez, I.; Masegosa, J.; Husillos, C.; del Olmo, A.; Perea, J.; Alfaro, E.J.; Fernández-Soto, A.; Moles, M.; Aguerri, J.A.L.; et al. Quasi-stellar objects in the ALHAMBRA survey. I. Photometric redshift accuracy based on 23 optical-NIR filter photometry. *Astron. Astrophys.* **2012**, *542*, A20. [CrossRef]
80. van Haarlem, M.P.; Wise, M.W.; Gunst, A.W.; Heald, G.; McKean, J.P.; Hessels, J.W.T.; de Bruyn, A.G.; Nijboer, R.; Swinbank, J.; Fallows, R.; et al. LOFAR: The LOw-Frequency ARray. *Astron. Astrophys.* **2013**, *556*, A2. [CrossRef]
81. Ochsenbein, F.; Bauer, P.; Marcout, J. The VizieR database of astronomical catalogues. *Astron. Astrophys.* **2000**, *143*, 23–32. [CrossRef]
82. Astropy Collaboration; Robitaille, T.P.; Tollerud, E.J.; Greenfield, P.; Droettboom, M.; Bray, E.; Aldcroft, T.; Davis, M.; Ginsburg, A.; Price-Whelan, A.M.; et al. Astropy: A community Python package for astronomy. *Astron. Astrophys.* **2013**, *558*, A33. [CrossRef]
83. Astropy Collaboration; Price-Whelan, A.M.; Sipőcz, B.M.; Günther, H.M.; Lim, P.L.; Crawford, S.M.; Conseil, S.; Shupe, D.L.; Craig, M.W.; Dencheva, N.; et al. The Astropy Project: Building an Open-science Project and Status of the v2.0 Core Package. *Astron. J.* **2018**, *156*, 123. [CrossRef]
84. Taylor, M.B. TOPCAT & STIL: Starlink Table/VOTable Processing Software. In Proceedings of the Astronomical Data Analysis Software and Systems XIV, Pasadena, CA, USA, 24–27 October 2004; Shopbell, P., Britton, M., Ebert, R., Eds.; Astronomical Society of the Pacific Conference Series; Astronomical Society of the Pacific: San Francisco, CA, USA, 2005; Volume 347, p. 29.