Radio Galaxy Detection Prediction with Ensemble Machine Learning

Rodrigo Carvajal, Israel Matute, José Afonso, Stergios Amarantidis and Davi Barbosa

Abstract The study of Active Galactic Nuclei (AGN) is fundamental to comprehend their evolution and connection with star-formation history and galaxy evolution. Powerful radio emission from AGN traces the largest structures in the Universe and, given the radiative and kinetic energy associated with the phenomena, is a prime feedback candidate to understand the correlations between the properties of the Super-massive Black Hole (SMBH) and the host galaxy. The Epoch of Reionisation (EoR) witnessed the birth of the first luminous sources and the initial steps of the SMBH-host connection. But few AGN have been identified in the EoR of which only a small fraction have radio detections. Recent and future large-scale surveys render the use of regular AGN detection and redshift estimation techniques inefficient. On the other hand, Machine Learning (ML) methodologies can help overcome the computational bottleneck to predict the presence of an ever-increasing number of AGN up to the highest redshifts.
We have developed a series of ML models that, using multi-band photometry, select a list of candidates with their predicted redshift. Models were trained and tested on the The Hobby-Eberly Telescope Dark Energy Experiment (HETDEX) and Stripe 82 fields. We find that AGN selection and redshift estimation metrics are similar to traditional techniques but with a fraction of the computational cost. The pipeline recovers 50-60% of the radio population, 2x-5x better than chance selection, allowing to shed some light on the origin and duty cycle of radio emission.

Rodrigo Carvajal · Israel Matute · José Afonso · Stergios Amarantidis · Davi Barbosa
Instituto de Astrofísica e Ciências do Espaço, Universidade de Lisboa, OAL, Tapada da Ajuda, PT1349-018 Lisbon, Portugal, e-mail: rcarvajal@alunos.fc.ul.pt

Rodrigo Carvajal · Israel Matute · José Afonso · Davi Barbosa
Departamento de Física, Faculdade de Ciências, Universidade de Lisboa, Edifício C8, Campo Grande, PT1749-016 Lisbon, Portugal

Stergios Amarantidis
Institut de Radioastronomie Millimétrique (IRAM), Avenida Divina Pastora 7, Local 20, 18012 Granada, Spain

1 Overview

Powerful radio emission from Active Galactic Nuclei (AGN) is a prime tracer of active accretion into a Super-massive Black Hole (SMBH). The reasons for its triggering (occurring for ∼5-30% of AGNs), are still highly uncertain. This radio emission is thought to be one of the mechanisms responsible for the observed correlation between the properties of the host galaxy and the SMBH. Confirming their nature requires detailed knowledge of their demographics up to the highest redshifts which we are still lacking.

In addition, the advent of recent and future surveys with next-generation observatories will render traditional analysis techniques difficult to apply at such large scale (e.g. Evolutionary Map of the Universe, EMU [15]). Advancing our knowledge of the field requires therefore alternative and efficient approaches to identify these sources at all redshift. Machine Learning (ML) provides the tools to understand very large volumes of data without extremely long computing time.

With this goal in mind, we have created a pipeline based on ML techniques to predict AGN at radio wavelengths and their redshift value using only photometric information from non-radio bands.

2 Methodology

We have gathered multi-wavelength photometric data for ∼6.7 million Wide-field Infrared Survey Explorer (WISE) [19] detected sources from the CatWISE2020 catalogue [14] over 424 $\deg^2$ in the Hobby-Eberly Telescope Dark Energy Experiment (HETDEX) Spring field [10]. These photometric data (cross-matched with Panoramic Survey Telescope and Rapid Response System [Pan-STARRS, 8] DR1, Two Micron All-Sky Survey [2MASS, 18, 3], and AllWISE [4] surveys) were used for training, testing and validation of the models and the prediction pipeline, On the other side, we have collected, in the same fashion as with HETDEX, multi-wavelength data from ∼370 000 CatWISE2020-detected sources from ∼92 $\deg^2$ in the Stripe 82 field (S82), which serves as an alternative testing field. Known spectroscopic labels (AGN-galaxy classification) and redshifts for the main catalogue are provided by the Million Quasar Catalog (MQC, [7]) and The Sloan Digital Sky Survey (SDSS)-DR16 [1] for 83 409 sources in the HETDEX field and for 3 304 sources in S82. The radio photometric information was obtained by cross-matching the main catalogue with several radio surveys [17, 5, 11, 9, 12] and a flag was created displaying whether or not a source shows a detection in, at least, one of the selected radio surveys. Only sources with spectroscopic classification and redshift were used for training, testing, and validation steps. A 20% of the spectroscopically-confirmed data was used as validation sub-set. From the remaining sources, 80% of them were used for training and 20% for testing.

The final pipeline is composed of three different instances executed in sequence: *i*) AGN-galaxy classification; *ii*) Radio detection prediction and *iii*) redshift estimate.

Field	F1	MCC	Recall
HETDEX	0.5255	0.4733	0.5235
Stripe 82	0.5395	0.3554	0.5386

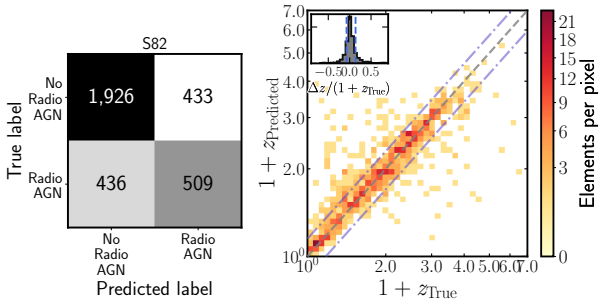

Fig. 1 Summary of results from predictions. *Left*) Overall metrics of the pipeline applied to the HETDEX and S82 fields: F1-score, Matthews Correlation Coefficient (MCC), and Recall. *Centre*) Confusion matrix for radio detection prediction on S82. *Right*) Comparison between true redshifts and those predicted by our pipeline over the sources predicted to be radio AGN in S82. Outliers lie outside the dashed-dotted line. Inset shows the dispersion between true and predicted values with a mean value of $\langle \frac{\Delta z}{1+z_{\rm True}} \rangle = 0.0768$.

Following the results presented by [2], we have used an ensemble of models for the training stages. Each instance finds the best models (meta learner) to stack the results from the remaining algorithms (base learners) as well as the best subset of features to use (from a total of 66 possibilities, including magnitudes as colours). For each prediction step, five tree-based models were tested: `CatBoost`, `XGBoost`, `ExtraTrees`, `Gradient-Boosting`, and `RandomForest`. For the AGN-galaxy classification and radio-detection models, the selected meta learner is `CatBoost`. In the case of the redshift value prediction, the meta model is `RandomForest`. The remaining models are used as base learners.

3 Results

Metrics based on the pipeline results (F1 score, Matthews Correlation Coefficient, and Recall), as well as the confusion matrix and true-predicted redshift comparison, are presented in Fig. 1. Despite the heterogeneity of the data, the AGN selection and redshift determination metrics are in line with recent results in the literature based in more homogeneous datasets. Our main finding is the significant improvement in the selection of radio AGN sources (more than 52% of recall) which is 2-5x better than a random selection of radio sources from an AGN sample (the fraction of radio-loud AGN is in the range of $5 - 30\%$) [16, 6, 13]. In addition, we find that the metrics of the pipeline are consistent when applied to a non-trained field (such as S82).

In addition, and when applied to non-labelled CatWISE2020 detections, the pipeline predicts an additional 99 723 and 7 047 new radio AGN candidates in HETDEX and S82 respectively.

Acknowledgements This work has been supported by the Fundação para a Ciência e a Tecnologia (FCT) through the Fellowship PD/BD/150455/2019 (PhD::SPACE Doctoral Network

PD/00040/2012) and POCH/FSE (EC) and through research grants PTDC/FIS-AST/29245/2017, UID/FIS/04434/2019, UIDB/04434/2020 and UIDP/04434/2020.

References

[1] Ahumada, R., Prieto, C.A., et al.: The 16th Data Release of the Sloan Digital Sky Surveys: First Release from the APOGEE-2 Southern Survey and Full Release of eBOSS Spectra. ApJS **249**(1), 3 (2020). DOI 10.3847/1538-4365/ab929e

[2] Carvajal, R., Matute, I., et al.: Exploring New Redshift Indicators for Radio-Powerful AGN. Galaxies **9**(4), 86 (2021). DOI 10.3390/galaxies9040086

[3] Cutri, R.M., Skrutskie, M.F., et al.: 2MASS All Sky Catalog of point sources. (2003)

[4] Cutri, R.M., Wright, E.L., et al.: Explanatory Supplement to the AllWISE Data Release Products (2013)

[5] de Gasperin, F., Williams, W.L., et al.: The LOFAR LBA Sky Survey. I. Survey description and preliminary data release. A&A **648**, A104 (2021). DOI 10.1051/0004-6361/202140316

[6] della Ceca, R., Lamorani, G., et al.: The Properties of X-Ray-selected Active Galactic Nuclei. III. The Radio-quiet versus Radio-loud Samples. ApJ **430**, 533 (1994). DOI 10.1086/174428

[7] Flesch, E.W.: The Million Quasars (Milliquas) v7.2 Catalogue, now with VLASS associations. The inclusion of SDSS-DR16Q quasars is detailed. arXiv e-prints arXiv:2105.12985 (2021)

[8] Flewelling, H.A., Magnier, E.A., et al.: The Pan-STARRS1 Database and Data Products. ApJS **251**(1), 7 (2020). DOI 10.3847/1538-4365/abb82d

[9] Gordon, Y.A., Boyce, M.M., et al.: A Catalog of Very Large Array Sky Survey Epoch 1 Quick Look Components, Sources, and Host Identifications. Research Notes of the American Astronomical Society **4**(10), 175 (2020). DOI 10.3847/2515-5172/abbe23

[10] Hill, G.J., Gebhardt, K., et al.: The Hobby-Eberly Telescope Dark Energy Experiment (HETDEX): Description and Early Pilot Survey Results. In: T. Kodama, T. Yamada, K. Aoki (eds.) Panoramic Views of Galaxy Formation and Evolution, *Astronomical Society of the Pacific Conference Series*, vol. 399, p. 115 (2008)

[11] Hodge, J.A., Becker, R.H., et al.: High-resolution Very Large Array Imaging of Sloan Digital Sky Survey Stripe 82 at 1.4 GHz. AJ **142**(1), 3 (2011). DOI 10.1088/0004-6256/142/1/3

[12] Intema, H.T., Jagannathan, et al.: The GMRT 150 MHz all-sky radio survey. First alternative data release TGSS ADR1. A&A **598**, A78 (2017). DOI 10.1051/0004-6361/201628536

[13] Macfarlane, C., Best, P.N., et al.: The radio loudness of SDSS quasars from the LOFAR Two-metre Sky Survey: ubiquitous jet activity and constraints on star formation. MNRAS **506**(4), 5888–5907 (2021). DOI 10.1093/mnras/stab1998

[14] Marocco, F., Eisenhardt, P.R.M., et al.: The CatWISE2020 Catalog. ApJS **253**(1), 8 (2021). DOI 10.3847/1538-4365/abd805

[15] Norris, R.P., Hopkins, A.M., et al.: EMU: Evolutionary Map of the Universe. PASA **28**(3), 215–248 (2011). DOI 10.1071/AS11021

[16] Padovani, P.: The radio-loud fraction and its dependence on magnitude and redshift. MNRAS **263**, 461–470 (1993). DOI 10.1093/mnras/263.2.461

[17] Shimwell, T.W., Tasse, C., et al.: The LOFAR Two-metre Sky Survey. II. First data release. A&A **622**, A1 (2019). DOI 10.1051/0004-6361/201833559

[18] Skrutskie, M.F., Cutri, R.M., et al.: The Two Micron All Sky Survey (2MASS). AJ **131**(2), 1163–1183 (2006). DOI 10.1086/498708

[19] Wright, E.L., Eisenhardt, P.R.M., et al.: The Wide-field Infrared Survey Explorer (WISE): Mission Description and Initial On-orbit Performance. AJ **140**(6), 1868–1881 (2010). DOI 10.1088/0004-6256/140/6/1868