Astronomy & Astrophysics manuscript no. ML_pipeline
July 19, 2023

Selection of Powerful Radio Galaxies with Machine Learning

R. Carvajal[1,2], I. Matute[1,2], J. Afonso[1,2], R. P. Norris[3,4], K. J. Luken[3,5], P. Sánchez-Sáez[6], P. Cunha[7,8], A. Humphrey[7,9], H. Messias[10,11], S. Amarantidis[12,1], D. Barbosa[1,2], H. A. Cruz[13], H. Miranda[1,2], A. Paulino-Afonso[7], and C. Pappalardo[1,2]

[1] Instituto de Astrofísica e Ciências do Espaço, Universidade de Lisboa, OAL, Tapada da Ajuda, 1349-018 Lisbon, Portugal
e-mail: racarvajal@ciencias.ulisboa.pt
[2] Departamento de Física, Faculdade de Ciências, Universidade de Lisboa, Edifício C8, Campo Grande, 1749-016 Lisbon, Portugal
[3] School of Science, Western Sydney University, Locked Bag 1797, Penrith, NSW 2751, Australia
[4] CSIRO Space & Astronomy, Australia Telescope National Facility, P.O. Box 76, Epping, NSW 1710, Australia
[5] CSIRO Data61, P.O. Box 76, Epping, NSW 1710, Australia
[6] European Southern Observatory, Karl-Schwarzschild-Straße 2, 85748 Garching bei München, Germany.
[7] Instituto de Astrofísica e Ciências do Espaço, Universidade do Porto, CAUP, Rua das Estrelas, 4150-762 Porto, Portugal
[8] Departamento de Física e Astronomia, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre 687, 4169-007 Porto, Portugal
[9] DTx - Digital Transformation CoLAB, Building 1, Azurém Campus, University of Minho, 4800-058 Guimarães, Portugal
[10] Joint ALMA Observatory, Alonso de Córdova 3107, Vitacura 763-0355, Santiago, Chile
[11] European Southern Observatory, Alonso de Córdova 3107, Vitacura, Casilla 19001, Santiago de Chile, Chile
[12] Institut de Radioastronomie Millimétrique (IRAM), Avenida Divina Pastora 7, Local 20, 18012 Granada, Spain
[13] Closer Consultoria Lda, Torre 1, Av. Eng. Duarte Pacheco 15, 1070-101 Lisboa, Portugal

Received ; accepted

ABSTRACT

Context. The study of Active Galactic Nuclei (AGN) is fundamental to discern formation and growth of supermassive black holes (SMBHs) and their connection with star-formation and galaxy evolution. **Due to the significant kinetic and radiative energy emitted by powerful AGN, they are prime candidates to observe the interplay between SMBH and stellar growth in galaxies.**
Aims. **We aim** to develop a method to predict the AGN nature of a source, its radio detectability and redshift **purely based on photometry**. The **use** of such method will increase the number of radio AGN, allowing us to improve our knowledge of accretion power into SMBH, the origin and triggers of radio emission and its impact on galaxy evolution.
Methods. We developed **and trained** a pipeline of three Machine Learning (ML) models **than can** predict **which sources are more likely to be** an AGN and to be detected in specific radio surveys. Also, it can estimate redshift values for predicted radio-detectable AGN. These models, which combine predictions from tree-based and gradient-boosting algorithms, have been trained with multi-wavelength data from near infrared-selected sources in The Hobby-Eberly Telescope Dark Energy Experiment (HETDEX) Spring Field. Training, testing, calibration and validation were carried out in the HETDEX field. Further validation was performed on near infrared-selected sources in the Stripe 82 field.
Results. In the HETDEX validation sub-set, our pipeline recovers 96% of the initially-labelled AGN and, from AGN candidates, we recover 50% of previously-detected radio sources. For Stripe 82, these numbers are 94% and 55%. **Compared to random selection, these rates are two and four times better for HETDEX, and** 1.2 **and** 12 **times better for Stripe 82, respectively. The pipeline can also recover the redshift distribution of these sources with** $\sigma_{\rm NMAD}$=0.07 **for HETDEX (**$\sigma_{\rm NMAD}$=0.09 **for Stripe 82) and an outlier fraction of** 19% **(**25% **for Stripe 82), compatible with previous results based on broadband photometry.** Feature importance analysis stresses the **relevance** of near- and mid-IR colours to select AGN and identifying their radio **and redshift** nature.
Conclusions. **Combining different algorithms** in ML models shows an improvement in the prediction power of our pipeline **over a random selection of sources**. Tree-based ML models (in contrast to Deep Learning techniques) facilitate the analysis of the impact that features have on the predictions. This prediction can give insight into the potential physical interplay between the properties of radio AGN (e.g. mass of black hole, accretion rate, etc).

Key words. Galaxies: active – Radio continuum: galaxies – Galaxies: high-redshift – Catalogs – Methods: statistical.

1. Introduction

Active Galactic Nuclei (AGN) are instrumental to determine the nature, growth, and evolution of supermassive black holes (SMBHs). Their strong emission allows us to study the close environment within the hosting galaxies and, at a larger scale, the intergalactic medium (e.g. Padovani et al. 2017; Bianchi et al. 2022). **Feedback due to AGN energetics, most prominently manifested in the form of jetted radio emission, might play a fundamental role in regulating stellar growth and overall evolution of hosts and their environments (Alatalo et al. 2015; Villar-Martín et al. 2017; Hardcastle & Croston 2020).**

Although radio emission can trace high star-formation in galaxies, above certain luminosities (e.g. log$L_{1.4\rm GHz}$>25 W Hz^{-1}, Jarvis et al. 2021) it is a prime tracer of the powerful jet emission **triggered by** the SMBH in AGN (Radio Galaxies, Heckman & Best 2014). Traditionally, these powerful Radio Galaxies (RGs) were used to pinpoint AGN activity but have been superseded in the last decades by optical, NIR, and X-ray surveys.

In fact, RGs in the high redshift Universe ($z > 2$) have been identified and studied **mostly** through the follow-up of AGN selected at shorter wavelengths (optical, NIR, millimetre, and X-rays, e.g. McGreer et al. 2006; Pensabene et al. 2020; Delhaize et al. 2021). The landscape is quickly changing and the advent of new radio instruments and surveys has allowed the detection of larger numbers of RGs (e.g. Williams et al. 2018; Capetti et al. 2020). Some of these surveys are: **the NRAO VLA Sky Survey (NVSS; Condon et al. 1998),** the Faint Images of the Radio Sky at Twenty-Centimetres (FIRST; Helfand et al. 2015), the Evolutionary Map of the Universe (EMU; Norris et al. 2011), the Very Large Array Sky Survey (VLASS; Gordon et al. 2020), and the LOFAR Two-metre Sky Survey (LoTSS; Shimwell et al. 2019).

One of the ultimate goals is to detect powerful RGs in the Epoch of Reionization (EoR) that could be used to trace the neutral gas distribution during this critical phase of the Universe (e.g. Carilli et al. 2004; Jensen et al. 2013). Simulations have shown that as much as a few hundreds of RGs per deg^2 could be present in the EoR (Amarantidis et al. 2019; Bonaldi et al. 2019; Thomas et al. 2021) and detectable with present and future deep observations –e.g. Square Kilometre Array (SKA), which is projected to have μJy **point-source** sensitivity levels **(SKA1-Mid is expected to reach close to** $2\,\mu$**Jy in** 1**-hour continuum observations at** $\nu \gtrsim 1$ **GHz; Prandoni & Seymour 2015; Braun et al. 2019)**–. **M**ost recent observational compilations (e.g. Inayoshi et al. 2020; Ross & Cross 2020; Bosman 2022; Fan et al. 2022), **show that** around 300 AGN have been confirmed to exist at redshifts higher than z∼6 over thousands of sq. degrees. This disagreement **highlights the uncertainties present in simulations, mainly due to our lack of knowledge of the triggering mechanisms and duty cycle for jetted emission in AGN (Afonso et al. 2015; Pierce et al. 2022).**

The selection of AGN candidates has had success in the X-rays and radio wavebands as they dominate the emission above certain luminosities. Unfortunately, deep X-ray surveys are limited in area and only of the order of 10% of AGN have strong radio emissions linked to jets **(i.e. radio-loud sources)** at any given time with variations, **going from** ∼ 6% up to ∼ 30%, correlated to optical and X-ray luminosities**, as well as with redshift** (e.g. Padovani 1993; della Ceca et al. 1994; Jiang et al. 2007; Storchi-Bergmann & Schnorr-Müller 2019; Gürkan et al. 2019; Macfarlane et al. 2021; Gloudemans et al. 2021, 2022; Best et al. 2023).[1]

The largest number of **AGN** candidates have been selected through the compilation of multi-wavelength energy distributions (SED) for millions of sources (Hickox & Alexander 2018; Pouliasis 2020). Of particular relevance for AGN are the **mid**-IR colours where *Spitzer* (Werner et al. 2004) and especially the Wide-field Infrared Survey Explorer (*WISE*; Wright et al. 2010) have opened a window for the detection of AGN over the whole sky, including the elusive fraction of heavily obscured ones (e.g. Stern et al. 2012; Mateos et al. 2012; Jarrett et al. 2017; Assef et al. 2018; Barrows et al. 2021).

Currently, extensive spectroscopic follow-up measurements have allowed the confirmation of the estimated redshifts for more than 800 000 AGN over large areas of the sky (Flesch 2021). Spectroscopic surveys have also contributed to the detection of AGN activity through the analysis of line ratio as is the case of the BPT diagram (Baldwin, Phillips, & Terlevich 1981). However, their determination can take long integration times and high-quality observations, rendering them not suited for **most sources in** large-sky catalogues. Photometric classification and redshifts (photo-z), are a viable option to understand the source nature and distribution across cosmic time (Baum 1957; Salvato et al. 2019). **Photometric** redshift estimations have been obtained for galaxies (e.g. Hernán-Caballero et al. 2021), and AGN (e.g. Ananna et al. 2017). **Template-fitting** photo-z estimations are computationally expensive **and require High-Performance Computing facilities** for large catalogues Gilda et al. ($\gtrsim 10^7$, 2021). **At the expense of redshift precision,** the use of drop-out techniques **offer a more computationally efficient solution** to generate and study high-redshift sources or candidates that, otherwise, would not have enough information to produce a precise redshift value (e.g. Bouwens et al. 2020; Carvajal et al. 2020; Shobhana et al. 2023).

Alternative statistical and computational methods can analyse **a large number** of elements and find relevant trends among their properties. One branch of these techniques **is Machine Learning (ML; Samuel 1959), which** can, using previously modelled data, predict the behaviour new data will have –i.e. the values of their properties–. In Astronomy, ML has been used with much success in a wide range of subjects, such as redshift determination, morphological classification, emission prediction, anomaly detection, observations planning, and more (e.g. Ball & Brunner 2010; Baron 2019). Traditional ML models are, in general, only fed with measurements and not with physical assumptions (Desai & Strachan 2021), and they do not need to check the consistency of the predictions or the results they provide. As a consequence, prediction times of traditional ML methods are typically less than **those from** physically-based methods.

Despite the large number of applications it might have, one important criticism that ML has received is related to the lack of interpretability –or explainability, as it is called in ML jargon– of the derived models, trends, and correlations. Most ML models, after taking a series of measurements and properties as input, deliver a prediction of a different property. But they cannot provide coefficients or an analytical expression, that might allow to find an equation for future predictions (Goebel et al. 2018). **An** important counter-example of this fact is the use of Symbolic Regression (e.g. Cranmer et al. 2020; Villaescusa-Navarro et al. 2021; Cranmer 2023). This implies that, for most ML models, it is not a simple task to understand which properties, and to what extent, help predict and interpret another attribute. This fact hinders our capability to understand the results in physical terms.

Recent work has been done to overcome the lack of explainability in ML models. The most widely used assessment is done with Feature Importance (Casalicchio et al. 2019; Roscher et al. 2020), both global and local (Saarela & Jauhiainen 2021). Game theory based analyses, **such as** the Shapley analysis (Shapley 1953), have also been used to understand **the** importance **of features** in Astrophysics (e.g. Machado Poletti Valle et al. 2021; Carvajal et al. 2021; Dey et al. 2022; Anbajagane et al. 2022; Alegre et al. 2022).

A further complication is that astronomical data can be very heterogeneous. Surveys and instruments gather data from many different areas in the sky with very different sensitivities and observational properties. This heterogeneity severely complicates most astronomical analyses, but in particular ML methods, as they are driven most of the time completely by the data. **This issue can be alleviated using observations in large, and homogeneous, surveys. Currently, among others, the Very Large Array (VLA), the Low Frequency Array (LOFAR), and the Giant Metrewave Radio Telescope (GMRT) allow obtaining**

[1] **Depending on the dataset, a random selection of AGN would lead to a rate of radio-detectable AGN in the range** 6 – 30%**. We will call this random choice a 'no-skill' selection.**

such measurements. Next-generation observatories and surveys –e.g. SKA, LSST, etc.– will also help in this regard, where observations will be carried out homogeneously over very large areas.

From a pure ML-based standpoint, several techniques used to lessen the effect of data heterogeneity have been developed (i.e. data cleansing and homogenisation). Some of them include discarding sources which add noise to the overall data distribution (Ilyas & Rekatsinas 2022). This can be extended to vetoing sources from specific areas in the sky (due to, for example, bad data reduction). Opposite to that, and when possible, previously mentioned techniques can be combined with increasing the survey area as a way to reduce possible biases. After selecting the data sample to be used for modelling, it is also possible to homogenise the measured ranges of observed properties. This procedure implies, for instance, normalising or standardising measured values can help ML models extract trends and connections among features more easily (Singh & Singh 2020).

Future observatories and surveys will deliver immense datasets. One option to analyse such observations **and confirming their radio AGN nature** is through **visual inspection (e.g. Banfield et al. 2015).** The use of such technique over large areas **can** have a very high cost. An alternative is using already-available multi-wavelength data **and template-fitting tools** to determine the likelihood of a**n AGN** to be detected in **radio** wavelength**s** (see, for instance, Pacifici et al. 2023). With the use of existing data, ML can help speeding this process up via the training of models that can detect **counterparts** in large radio surveys (see, for an example of the efforts done to achieve this goal, Hopkins et al. 2015; Bonaldi et al. 2021).

Building upon the work presented by Carvajal et al. (2021), we aim to identify candidates of high-redshift radio-detectable AGN which can be extracted from heterogeneous large-area surveys. We **have** developed a series of ML models to predict, separately, the detection of AGN, the detection of the radio signal from AGN, and the redshift values of radio-detectable AGN **using non-radio photometric data. In this way, it might be possible to avoid the direct analysis of large numbers of radio detections**. Furthermore, we tested the performance of these models without applying a large number of previous cleaning steps, which **might** reduce considerably the size of the training sets. The compiled catalogue of candidates can help using data from future large-sky surveys more efficiently, as observational and analytical efforts can be focused on the areas in which AGN have been predicted to exist.

We seek**, therefore,** to test **the generalisation power of** such models by applying them in **a different area** from the training field with data that is not necessarily of the same quality.

The structure of this article is as follows. In Sect. 2, we present the data and its preparation for ML training. The selection of models and the metrics used to assess their results are shown in Sect. 3. In Sect. 4, the results of model training and validation are shown as well as the predictions using the ML pipeline for radio AGN detections. We present the discussion of our results in Sect. 5. Finally, in Sect. 6, we summarise our work.

2. Data

A large area with deep and homogeneous quality radio observations is needed to train and validate our models and predictions for RGs with already existent observations. As training field we selected the area of the Hobby-Eberly Telescope Dark Energy Experiment Spring Field (HETDEX; Hill et al. 2008) covered by the first data release of the LOFAR Two-metre Sky Survey (LoTSS-DR1; Shimwell et al. 2019). The LoTSS-DR1 survey covers 424 deg^2 in the HETDEX Spring field (hereafter, HETDEX field) with LOFAR (van Haarlem et al. 2013) 150 MHz observations that have a median sensitivity of 71 μJy/beam **and an angular resolution of** 6″. HETDEX provides as well multi-wavelength homogeneous coverage as described below.

In order to test the performance of the models when applied to different areas of the sky, and with different coverages from radio surveys, we have selected the Sloan Digital Sky Survey (SDSS, York et al. 2000) Stripe 82 Field (S82, Annis et al. 2014; Jiang et al. 2014). For S82, we collected data from the same surveys as with the HETDEX (see the following section) field but with one important caveat: no LoTSS-DR1 data is available in the field and, thus, we gathered the radio information from the VLA SDSS Stripe 82 Survey (VLAS82; Hodge et al. 2011). VLAS82 covers an area of 92 deg^2 with a median rms noise of 52 μJy/beam at 1.4 GHz **with an angular resolution of** 1.″8. We have selected the **S82** field (and, in particular, the area covered by VLAS82) given that it presents deep radio observations but taken with a different instrument than LOFAR. This difference allows us to test the suitability of our models and procedures in conditions that are different from the training circumstances.

2.1. Data collection

The base survey from which all the studied sources have been drawn is the CatWISE2020 catalogue (CW; Marocco et al. 2021). It lists NIR-detected elements selected from WISE (Wright et al. 2010) and NEOWISE (Mainzer et al. 2011, 2014) over the entire sky at 3.4 and 4.6 μm (W1 and W2 bands, respectively). This catalogue includes sources detected at 5σ in either of the used bands (i.e. W1∼17.43 and W2∼16.47 mag$_{\rm Vega}$ respectively). The HETDEX field contains 15 136 878 sources listed in CW. Conversely, in the **S82** field, there are 3 590 306 of them.

Multi-wavelength counterparts for CW sources were found on other catalogues applying a 5″search criteria. These catalogues include Pan-STARRS DR1 (PS1; Chambers et al. 2016; Flewelling et al. 2020), 2MASS All-Sky (2M; Skrutskie et al. 2006; Cutri et al. 2003a,b), and AllWISE (AW; Cutri et al. 2013)[2]. The adopted search radius corresponds to the distance that has been used by Wright et al. (2010) to match radio sources to Pan-STARRS and WISE observations. Nevertheless, the source density of the radio (LOFAR and VLA) and 2MASS catalogues imply a low statistical (< 1%) spurious counterpart association, this is not the case for PS1, where the source density is higher. For this reason, and to mantain a statistically low spurious association between CW and PS1, we limited our search radius to 1.″1. This distance corresponds to the smallest Point-Spread Function (PSF) size of the bands included in PS1 (Chambers et al. 2016).

For the purposes of this work, observations in LoTSS and VLAS82 are only used to determine whether a source is radio detected, or not. In particular, no check has been performed on whether a selected source is extended or not in any of the radio surveys. A single boolean feature is created from the radio measurements (see Sect. 2.2) and no further analyses were performed regarding the detection levels that might be found.

Additionally, we have discarded the measurement errors of all bands. Traditionally, ML algorithms cannot incorporate un-

[2] For the purposes of the analyses, and except when clearly stated otherwise, photometric measurements **are** converted to AB magnitudes.

Table 1. Bands available for model training in our dataset

Survey	Band (Column name)[a]
Pan-STARRS (PS1)	g (gmag), r (rmag), i (imag), z (zmag), y (ymag)
2MASS (2M)	J (Jmag), H (Hmag), Ks (Kmag)
CatWISE2020 (CW)	W1 (W1mproPM), W2 (W2mproPM)
AllWISE (AW)	W3 (W3mag), W4 (W4mag)

Notes. [(a)] In parentheses are shown the names of the columns or features in our dataset that represent each band.

certainties in a straightforward way and, thus, we opted to avoid attempting to use them for training**. One significant counter-example corresponds to Gaussian Processes (GPs; Rasmussen & Williams 2006), where measurement uncertainties are needed by the algorithm to generate predictions. Additionally, the astronomical community has attempted to modify existing techniques to include uncertainties in their ML studies. Some examples include the works by** Ball et al. (2008); Reis et al. (2019); Shy et al. (2022). Furthermore, Euclid Collaboration et al. (2023b) have shown that, in specific cases, the inclusion of measurement errors does not add new information to the training of the models and can be even detrimental to the prediction metrics. The degradation of the model by including uncertainties can likely be related to the fact that, by virtue of the large number of sources included in the training stages, the uncertainties are already encoded in the dataset in the form of scatter.

Following the same argument of measurement errors, upper limit values have been removed and a missing value is assumed instead. In general, ML methods (and their underlying statistical methods) cannot work with catalogues that have empty entries (Allison 2001). For that reason, we have used single imputation (a review on the use of this method**, which is part of data cleansing,** in astronomy can be seen in Chattopadhyay 2017) to replace these missing values, and those fainter than 5−σ limits, with meaningful quantities that represent the lack of a measurement. We have opted for the inclusion of the same 5−σ limiting magnitudes as the value to impute with. This method of imputation with some variations, has been successfully applied and tested, recently, by Arsioli & Dedin (2020); Carvajal et al. (2021); Curran (2022), and Curran et al. (2022). **In particular, Curran (2022) tested several data imputation methods. Among those which replaced all missing values in a wavelength band with a single, constant value, using the** 5−σ **limiting magnitudes showed the best performance.**

In this way, observations from 12 non-radio bands were gathered (as listed in Table 1). The magnitude density distribution for the sample from the HETDEX and S82 fields, without any imputation, is shown in Fig. 1. After imputation, the distribution of magnitudes changes, as shown in Fig. 2. Each panel of the figure shows the number of sources which have a measurement above its 5−σ limit in such band. Additionally, a representation of the observational 5−σ limits of the bands and surveys used in this work is presented in Fig. 3. It is worth noting the depth difference between VLAS82 and LoTSS-DR1 is ∼1.5 mag for a typical synchrotron emitting source ($F_\nu \propto \nu^\alpha$ with $\alpha = -0.8$), allowing the latter survey reach fainter sources.

AGN labels and redshift information were obtained by cross-matching (with a 1.″1 search radius) the catalogue with the Million Quasar Catalog[3] (MQC, v7.4d; Flesch 2021), which lists

[3] http://quasars.org/milliquas.htm

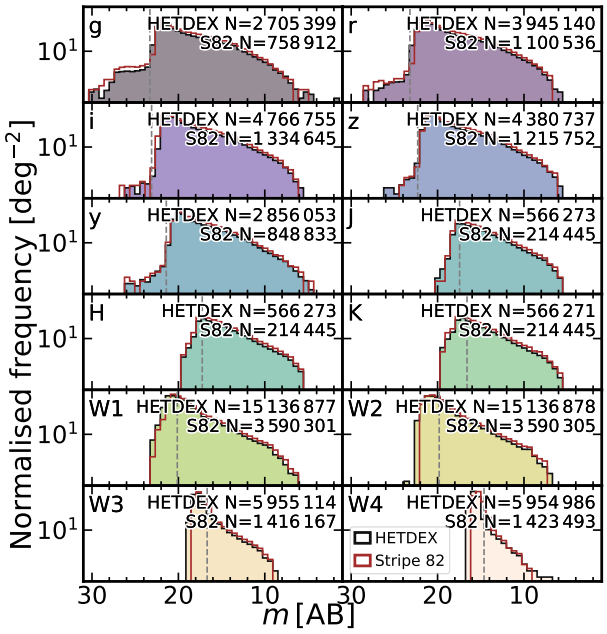


Fig. 1. Histograms of base collected, non-imputed, non-radio bands for HETDEX (clean, background histograms) and **S82** (empty, brown histograms). Each panel shows the distribution of measured magnitudes of detected sources divided by the total area of the field. Dashed, vertical lines represent the 5−σ magnitude limit for each band. The number in the upper right corner of each panel shows the number of measured magnitudes included in their corresponding histogram.

information from more than 1 500 000 objects that have been classified as optical QSO, AGN, or Blazars. Sources listed in the MQC may have additional counterpart information, including radio or X-ray associations. For the purposes of this work, only sources with secure spectroscopic redshifts were used. The matching yielded 50 538 spectroscopically confirmed AGN in HETDEX and 17 743 confirmed AGN in **S82**.

Similarly, the sources in our parent catalogue were cross-matched with the Sloan Digital Sky Survey Data Release 16 (SDSS-DR16; Ahumada et al. 2020). This cross-match was done solely to determine which sources have been spectroscopically classified as galaxies (`spClass == GALAXY`). For most of these galaxies, SDSS-DR16 lists a spectroscopic redshift value, which will be used in some stages of this work. In the HETDEX field, SDSS-DR16 provides 68 196 spectroscopically confirmed galaxies. In the **S82** field, SDSS-DR16 identifies 4 085 galaxies spectroscopically. Given that MQC has access to more AGN detection methods than SDSS, when sources were identified as both galaxies (in SDSS-DR16) and AGN (in the MQC), a final label of AGN was given. A description of the number of elements in each field and the multi-wavelength counterparts found for them is presented in Table 2.

From Table 2, it is possible to see that the numbers and ratios of AGN and galaxies in both fields are dissimilar. S82 has been subject to a larger number of observations, which have allowed the detection of a larger fraction of AGN than in the HETDEX field (see, for instance, Lyke et al. 2020), which does not have such number of dedicated studies.

Attending to the intrinsic differences between ML algorithms, not all of them have the same performance when being trained with features spanning a wide range of values (i.e. several orders of magnitude). For this reason, it is customary to re-scale the available values to either be contained within the range [0, 1]

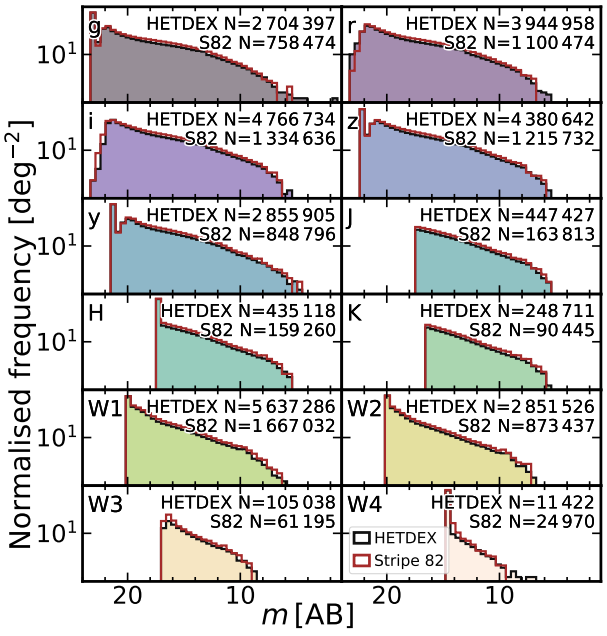


Fig. 2. Histograms of base collected non-radio bands for HETDEX (clean, background histograms) and **S82** (empty, brown histograms) fields. Description as in Fig. 1. The number in the upper right corner of each panel shows number of sources with magnitudes originally measured above the 5−σ limit included in their corresponding histogram for each field **(i.e. sources that have not been imputed or replaced)**.

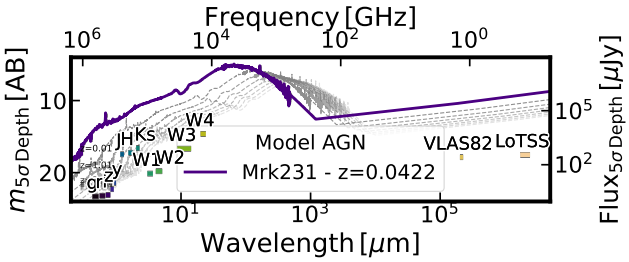


Fig. 3. Flux and magnitude depths (5−σ) from the surveys and bands used in this work. Limiting magnitudes and fluxes were obtained from the description of the surveys, as referenced in Sect. 2.1. In purple, rest-frame SED from Mrk231 (z = 0.0422, Brown et al. 2019) is displayed as an example AGN. Redshifted (from z=0.001 to z=7) versions of this SED are shown in dashed grey lines.

Table 2. Composition of initial catalogue and number of cross matches with additional surveys and catalogues.

Survey	HETDEX	**S82**
CatWISE2020	15 136 878	3 590 306
AllWISE	5 955 123	1 424 576
Pan-STARRS	4 837 580	1 346 915
2MASS	566 273	214 445
LoTSS	187 573	...
VLAS82	...	8 747
MQC (AGN)	50 538	17 743
SDSS (Galaxy)	68 196	4 085

or to have similar distributions. We applied a version of the latter transformation to our features (not the targets) as to have a mean value of $\mu = 0$ and a standard deviation of $\sigma = 1$ for each feature. Additionally, these new values were power-transformed to resemble a Gaussian distribution. This transformation helps the models avoid using the distribution of values as additional information for the training. For this work, a Yeo-Johnson transformation (Yeo & Johnson 2000) was applied.

2.2. Feature pool

The initial pool of features that have been selected or engineered to use in our analysis is briefly described below:

- Photometry, both measured and imputed, in the form of AB magnitudes for a total of 12 bands.
- Colours. All available colours from measured and imputed magnitudes were considered. In total, there are 66 colours, resulting from all available combinations of two magnitudes between the 12 selected bands. These colours are labelled in the form `X_Y` where `X` and `Y` are the respective magnitudes.
- Number of non-radio bands in which a source has valid measurements (`band_num`) . This feature could be, very loosely, attributed to the total flux a source can display. A higher `band_num` will imply that such source can be detected in more bands, hinting a higher flux (regardless of redshift). The use of features with counting or aggregation of elements in the studied dataset is well established in ML (see, for example, Zheng & Casari 2018; Duboue 2020; Sánchez-Sáez et al. 2021; Euclid Collaboration et al. 2023b).
- AGN-Galaxy classification boolean flag named `class`.
- Radio boolean flag `LOFAR_detect`. This feature flags whether sources have counterparts in the radio catalogues (LoTSS or VLAS82).

A list of the features created for this work and their representation in the code and in some of the figures is presented in Table A.1.

3. Machine Learning training

In an attempt to extract the largest available amount of information from the data, and let ML algorithms improve their predictions, we have decided to perform our training and predictions through a series of sequential steps, which we refer to as 'models' henceforth. We have started with the training and prediction of the class of sources (AGN or galaxies). The next model predicts whether an AGN could be detected in radio at the depth used during training (LoTSS). A final model will predict the redshift values of radio-predicted AGN. A visual representation of this process can be seen in Fig. 4. Creating separate models gives us the opportunity to select the best subset of features for training as well as the best combination of ML algorithms for training in each step.

In broad terms, our goal with the classification models is to recover the largest number of elements from the positive classes (`class = 1` and `LOFAR_detect = 1`). For the regression model, we aim to retrieve predictions as close as the originally fed redshift values.

In general, classification models provide a final score in the range [0, 1], which can only be associated with a true probability after a careful calibration (Kull et al. 2017a,b). Calibration of these scores can be done by applying a transformation to their values. For our work, we will apply a Beta transformation[4]. This type of transformation allows to re-distribute the scores of an

[4] Beta transformation functions have the general form $\mu_{beta}(S; a, b, c) = 1/\left(1 + 1/\left(e^{c}\frac{S^{a}}{(1-S)^{b}}\right)\right)$, with S being the score from the classifier and a, b, c, free parameters to be optimised.

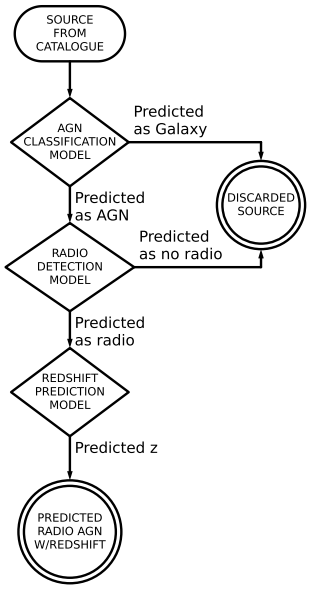
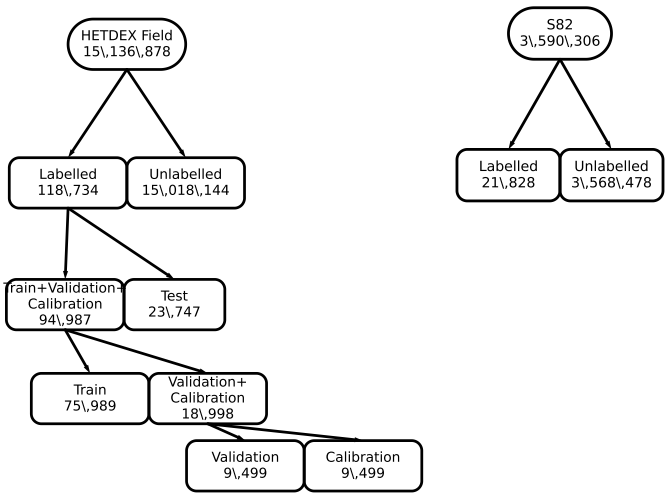


Fig. 4. Flowchart representing the prediction pipeline used to predict the presence of radio-detectable AGN and their redshift values. At the beginning of each model step, the most relevant features are selected as described in Sect. 3.1.



(a) HETDEX Field (b) **S82** Field

Fig. 5. Composition of datasets used for the different steps of this work. (a) HETDEX Field. (b) **S82**.

uncalibrated classifier allowing them to get closer to the definition of probability. Further details of this calibration are given in the Appendix C.

Given that we need to be able to compare the results from the training and application of the ML models with values obtained independently (i.e. ground truth), we divided our dataset into labelled and unlabelled sources. Labelled sources are all elements of our catalogue that have been classified as either AGN or galaxies. Unlabelled sources are those which lack such classification and that will **only** be subject to the prediction of our models**, not taking part in any training step.**

Before any calculation or transformation is applied to the data from the HETDEX field, we split the labelled dataset into training, validation, calibration, and testing subsets. The early creation of these subsets helps avoid information leakage from the test subset into the models. Initially, a 20% of the dataset has been reserved as testing data. Of the remaining elements, an 80% of them have been used for training, and the rest of the data has been divided equally between calibration and validation subsets (i.e. a 10% each). The splitting process and the number of elements for each subset are shown in Fig. 5. Depending on the model, the needed sources are selected from each of the sub-sets that have been already created. **The training set will be used to select algorithms for each step and to optimise their hyperparameters. The inclusion of the validation sub-set helps in the parameter optimisation of the models. The probability calibration of the trained model is performed over the calibration sub-set and, finally, the completed models are tested on the test sub-set.** The use of these subsets will be **expanded** in Sects. 3.3 and 3.4.

All the following transformations (feature selection, standardisation, and power transform of features) have been applied to the training and validation subsets before the training of the algorithms and models. The calibration and testing subsets were subject to the same transformations after the modelling stage.

3.1. Feature selection

ML algorithms, as with most data analysis tools, require execution times which increase at least linearly with the size of the datasets. In order to reduce training times without losing relevant information for the model, the most important features were selected at each step through a process called feature selection.

To avoid redundancy, the process starts discarding features that have a high correlation with another property of the dataset. For discarding features, we calculated Pearson's correlation matrix for the full train+validation dataset only and selected the pairs of features that showed a correlation factor higher than $\rho = 0.75$, in absolute values[5]. From each pair, we discarded the feature with the lowest relative standard deviation (RSD; Johnson & Leone 1964). The RSD is defined as the ratio between the standard deviation of a set and its mean value. A feature which covers a small portion of its probable values (i.e. low coverage of parameter space, and lower RSD) will give less information to a model than one with largely spread values.

For each model, the process of feature selection begins with 79 base features and three targets (`class`, `LOFAR_detect`, and `Z`). Feature selection is run, independently, for each trained model (i.e. AGN-Galaxy classification, radio detection, and redshift predictions), delivering three different sets of features.

3.2. Metrics

A set of metrics will be used to understand the reliability of the results and put them in context with results in the literature. Since our work includes the use of classification and regression models, we briefly discuss the appropriate metrics in the following sections.

3.2.1. Classification metrics

The main tool to assess the performance of classification methods is the Confusion (or Error) Matrix. It is a two-dimension

[5] A value of $\rho = 0.75$ is a compromise between very stringent thresholds (e.g. $\rho = 0.5$) and more relaxed values (e.g. $\rho \approx 0.9$). For an explanation on how to consider different correlation values, see, for instance Ratner (2009)

(predicted vs. true) matrix where the true and predicted class(es) are compared and results stored in cells with the rate of True Positives (TP), True Negative (TN), False Positives (FP), and False Negatives (FN). As mentioned earlier in Sect. 3, we seek to maximise the number of positive-class sources that are recovered as such. Using the elements of the confusion matrix, this aim can be translated into the maximisation of TP and, consequently, the minimisation of FN.

From the elements of the confusion matrix, we can obtain additional metrics, such as the F1 and F_β scores (Dice 1945; Sørenson 1948; van Rijsbergen 1979), and the Matthews Correlation Coefficient (MCC; Yule 1912; Cramér 1946; Matthews 1975) which are better suited for unbalanced data as they take into account the behaviour and correlations among all elements of the confusion matrix. As such, the F1 coefficient is defined as:

$$\mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FN} + \mathrm{FP}}\,. \tag{1}$$

F1 values can go from 0 (no prediction of positive instances) to 1 (perfect prediction of elements with positive labels). This definition assigns equal weight (importance) to both the number of FN and FP. An extension to the F1 score, which adds a non-negative parameter, β, to increase the importance given to each one of them is the F-Score (F_β), defined as:

$$\mathrm{F}_\beta = \frac{(1+\beta^2)\times \mathrm{TP}}{(1+\beta^2)\times \mathrm{TP} + \beta^2 \times \mathrm{FN} + \mathrm{FP}}\,. \tag{2}$$

Using $\beta > 1$, more relevance is given to the optimisation of FN. When $0 \leq \beta < 1$, the optimisation of FP is more relevant. If $\beta = 1$, the initial definition of F1 is recovered. As with F1, F_β values can be in the range $[0, 1]$. As we seek to minimise the number of FN detection, we adopt a conservative value of $\beta = 1.1$, giving more significance to their reduction without removing the aim for FP. Also, this value is close enough to $\beta = 1$, which will allow us to compare our scores to those produced in previous works.

MCC is defined as:

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN} - \mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}\,, \tag{3}$$

which includes also the information about the TN elements. MCC can range from -1 (total disagreement between true and predicted values) to $+1$ (perfect prediction) with 0 representing a prediction analogous to a random guess.

The Recall (also called Completeness, Sensitivity, or True Positive Rate -TPR-; Yerushalmy 1947) corresponds to the rate of relevant, or correct, elements that have been recovered by a process. Using the elements from the confusion matrix, it can be defined as:

$$\mathrm{Recall} = \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}. \tag{4}$$

The TPR can go from 0 to 1, with a value of 1 meaning that the model can recover all the true instances.

The last metric used is Precision (also known as Purity), which can be defined as the ratio between the number of correctly classified elements and the number of sources in the positive class (AGN or radio detectable):

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\,. \tag{5}$$

Precision can range from 0 to 1 where higher values show that more real positive instances of the studied set were retrieved as such by the model.

In order to establish a baseline from which the aforementioned metrics can be assessed, it is possible to obtain them in the case of a random, or no-skill prediction. Following, for instance, the derivations and notation from Poisot (2023), no-skill versions of classification metrics (Eqs. 2–5) are:

$$\mathrm{F}_\beta^{\mathrm{no-skill}} = p\,, \tag{6}$$

$$\mathrm{MCC}^{\mathrm{no-skill}} = 0\,, \tag{7}$$

$$\mathrm{Recall}^{\mathrm{no-skill}} = p\,, \tag{8}$$

$$\mathrm{Precision}^{\mathrm{no-skill}} = p\,. \tag{9}$$

where p corresponds to the ratio between the elements of the positive class and the total number of elements involved in the prediction.

3.2.2. Regression metrics

For the case of individual redshift value determination, two commonly used metrics are the difference between predicted and true redshift,

$$\Delta z = z_{\mathrm{True}} - z_{\mathrm{Predicted}}\,, \tag{10}$$

and its normalised difference,

$$\Delta z^{\mathrm{N}} = \frac{z_{\mathrm{True}} - z_{\mathrm{Predicted}}}{1 + z_{\mathrm{True}}}\,. \tag{11}$$

If the comparison is made over a larger sample of elements, the bias of the redshift is used (Dahlen et al. 2013), with the median of the quantities instead of its mean to avoid the strong influence of extreme values:

$$\Delta z_{\mathrm{Total}} = \mathrm{median}\,(z_{\mathrm{True}} - z_{\mathrm{Predicted}}) = \mathrm{median}(\Delta z)\,, \tag{12}$$

$$\Delta z^{\mathrm{N}}_{\mathrm{Total}} = \mathrm{median}\left(\frac{z_{\mathrm{True}} - z_{\mathrm{Predicted}}}{1 + z_{\mathrm{True}}}\right) = \mathrm{median}(\Delta z^{\mathrm{N}})\,. \tag{13}$$

Using the previous definitions, four additional metrics can be calculated. These are the median absolute deviation (MAD, σ_{MAD}) and normalised median absolute deviation (NMAD, σ_{NMAD}; Hoaglin et al. 1983; Ilbert et al. 2009), which are less sensitive to outliers. Also, the standard deviation of the predictions, σ_z, and its normalised version, σ_z^{N} are typically used. They are defined as:

$$\sigma_{\mathrm{MAD}} = 1.48 \times \mathrm{median}\,(|\Delta z|)\,, \tag{14}$$

$$\sigma_{\mathrm{NMAD}} = 1.48 \times \mathrm{median}\left(\left|\Delta z^{\mathrm{N}}\right|\right)\,, \tag{15}$$

$$\sigma_z = \sqrt{\frac{1}{\mathrm{d}}\sum_i^{\mathrm{d}} (\Delta z)^2}\,, \tag{16}$$

$$\sigma_z^{\mathrm{N}} = \sqrt{\frac{1}{\mathrm{d}}\sum_i^{\mathrm{d}} (\Delta z^{\mathrm{N}})^2}\,, \tag{17}$$

with d being the number of elements in the studied sample (i.e. its size).

Also, the outlier fraction (η, as used in Dahlen et al. 2013; Lima et al. 2022) is considered, which is defined as the fraction sources with a predicted redshift difference ($\left|\Delta z^{\rm N}\right|$, Eq. 11) larger than a previously set value. Taking the results from Ilbert et al. (2009) and Hildebrandt et al. (2010), we have selected this threshold to be 0.15, leaving the definition of the outlier fraction as:

$$\eta = \frac{\#\left(\left|\Delta z^{\rm N}\right| > 0.15\right)}{d} \,. \tag{18}$$

where # symbolises the number of sources fulfilling the described relation, and d corresponds to the size of the selected sample.

3.2.3. Calibration metrics

One of the most used analytical metrics to assess calibration of a model is the Brier score (BS; Brier 1950). It measures the mean square difference between the predicted probability of an element and its true class. If the total number of elements in the studied sample is d, the BS can be written (for binary classification problems, as the ones studied in this work) as:

$$\mathrm{BS} = \frac{1}{d}\sum_{i}^{d}(\mathbb{C} - \mathrm{class})^2 \,, \tag{19}$$

where $\mathbb{C}$ is the predicted class and class the true class of each of the elements in the sample (0 or 1). The BS can range between 0 and 1 with 0 representing a model that is completely reliable in its predictions.

Additionally, the BS can be used to compare the reliability (or calibration) between a model and a reference using the Brier Skill Score (BSS; e.g. Glahn & Jorgensen 1970):

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS_{ref}}} \,. \tag{20}$$

In our case, $\mathrm{BS_{ref}}$ corresponds to the value calculated from the uncalibrated model. The BSS can take values between -1 and $+1$. The closer the BSS gets to 1, the more reliable the analysed model is. These values include the case where BSS$\approx$0, in which both models perform similarly in terms of calibration.

For our pipeline, after a model has been fully trained, a calibrated version of their scores will be obtained. With both of them, the BSS will be calculated and, if it is not much lower than 0, that calibrated transformation will be used as the final scores from the prediction.

3.3. Model selection

By design, each ML algorithm has been developed and tuned to work better with certain data conditions, i.e. balance of target categories, ranges of base features, etc. The predicting power of different algorithms can be combined with the use of meta-learners (Vanschoren 2019). Meta-learners use the properties or predictions from other algorithms (base learners) as additional information during their training stages. A simple implementation of this procedure is called Generalised Stacking (Wolpert 1992) which can be interpreted as the addition of priors to the model training stage. Generalised stacking has been applied in several astrophysical problems. That is the case of Zitlau et al. (2016), Cunha & Humphrey (2022), and Euclid Collaboration et al. (2023a), Euclid Collaboration et al. (2023b).

Base and meta learners have been selected based upon the metrics described in Sect. 3.2. We have trained five algorithms with the training subset and calculated the metrics for all of them using a 10-fold cross-validation approach (e.g. Stone 1974; Allen 1974) over the same training subset. For each metric, the learners have been given a rank (from 1 to 5) and a mean value has been obtained from them. Out of the analysed algorithms, the one with the best overall performance (i.e. best mean rank) is selected to be the meta learner while the remaining four are used as base learners.

For the AGN-galaxy classification and radio detection problems, we tested five classification algorithms: Random Forest (`RF`; Breiman 2001), Gradient Boosting Classifier (`GBC`; Friedman 2001), Extra Trees (`ET`; Geurts et al. 2006), Extreme Gradient Boosting (`XGBoost`, `v1.5.1`; Chen & Guestrin 2016), and `CatBoost` (`v1.0.5`; Dorogush et al. 2017, 2018). For the redshift prediction problem, we tested five regressors as well: `RF`, `ET`, `XGBoost`, `CatBoost`, and Gradient Boosting Regressor (`GBR`; Friedman 2001). We have used the Python implementations of these algorithms and, in particular for `RF`, `ET`, `GBC`, and `GBR`, the versions offered by the package `scikit-learn`[6] (`v0.23.2`; Pedregosa et al. 2011). These algorithms were selected given that they offer tools to interpret the global and local influence of the input features in the training and predictions (cf. Sects. 1 and 5.3).

All the algorithms selected for this work fall into the broad family of Tree-Based models. Forest models (`RF` and `ET`) rely on a collection of decision trees to, after applying a majority vote, predict either a class or a continuum value. Each of these decision trees uses a different, randomly-selected sub-set of features to make a decision on the training set (Breiman 2001). Opposite to forests, Gradient Boosting models (`GBC`, `GBR`, `XGBoost` and `CatBoost`) apply decision trees sequentially to improve the quality of the previous predictions (Friedman 2001, 2002).

3.4. Training of models

The procedure described in Sect. 3.3 includes an initial fit of the selected algorithms to the training data (including the selected features) to optimise their parameters. The stacking step includes a new optimisation of the parameters of the meta-learner using 10-fold cross-validation on the training data with the addition of the output from the base learners, which are treated as regular features. Then, the hyper-parameters of the stacked models are optimised over the training sub-set (a brief description of this step is presented in Appendix D).

The final step involves a last parameter fitting instance but using, this time, the combined train+validation subset, which includes the output of the base algorithms, to ensure wider coverage of the parameter space and better-performing models. Consequently, only the testing set is available for assessing the quality of the predictions made by the models.

3.5. Probability calibration

The calibration procedure was performed in the calibration subset. In this way, we avoid influencing the process with information from the training and validation steps. A broader description

[6] `https://scikit-learn.org`

of the calibration process and the results obtained for our models are presented in Appendix C.

From this point onward, and with the sole exception of some of the outcomes shown in Sect. 5.3, all results from classifications will be based on the calibrated probabilities.

3.6. Optimisation of classification thresholds

As mentioned in the first paragraphs of Sect. 3, classification models deliver a range of probabilities for which a threshold is needed to separate their predictions between negative and positive classes. By default, these models set a threshold at 0.5 in score[7] but, in principle, and given the characteristics of the problem, a different optimal threshold might be needed.

In our case, we want to optimise (increase) the number of recovered elements in each model (i.e. AGN or radio-detectable sources). This maximisation corresponds to obtaining thresholds that optimise the recall given a specific precision limit. We did that with the use of the statistical tool called Precision-Recall (PR) Curve. A deeper description of this method and the results obtained from our work are presented in Appendix E[8].

4. Results

In the present section, we report the results from the training of the different models in the HETDEX field. All metrics are evaluated using the testing subset. The metrics are also computed on labelled AGN in the **S82** field. As no training is done on **S82** data, it offers a way to test the validity of the pipeline on data that, despite having similar optical-NIR photometric properties, presents distinct radio information and location in the sky.

The three models are chained afterwards in sequential mode to create a pipeline, and related metrics, for the prediction of radio-AGN activity. Novel predictions were obtained from the application of such pipeline to unlabelled sources from both the HETDEX and **S82** fields.

4.1. AGN-Galaxy classification

Feature selection was applied to the train+validation subset with 85 488 confirmed elements (galaxies from SDSS DR16 and AGN from MQC, i.e. `class == 0` or `class == 1`). After the selection procedure described in Sect. 3.1, 18 features were selected for training: `band_num`, `W4mag`, `g_r`, `r_i`, `r_J`, `i_z`, `i_y`, `z_y`, `z_W2`, `y_J`, `y_W1`, `y_W2`, `J_H`, `H_K`, `H_W3`, `W1_W2`, `W1_W3`, and `W3_W4`. The target feature is `class`.

The results of model testing for the AGN-galaxy classification are reported in Table 3. The `CatBoost` algorithm provides the best metric values (highest mean rank) and is therefore selected as the meta-model. `XGBoost`, `RF`, `ET`, and `GBC` were used as base learners.

The optimisation of the PR curve for the calibrated predictor provides an optimal threshold for this algorithm of 0.34895. This value was used for the AGN-Galaxy model throughout this work.

The results of the application of the stacked and calibrated model for the testing subset and the labelled sources in S82 are presented in Table 4. The metrics are shown for the use of two different thresholds, the naive value of 0.5 and the PR-derived value of 0.34895. The confusion matrix (calculated on the testing dataset) is shown in the upper left panel of Fig. 6.

[7] Throughout this work, we will call this a naive threshold.
[8] Thresholds derived from the PR curves will be labelled as PR.

Table 3. Best performing models for the AGN-galaxy classification

Model	F_β (×100)	MCC (×100)	Precision (×100)	Recall (×100)	Rank
`CatBoost`	95.70±0.28	92.46±0.48	95.45±0.32	95.91±0.37	1.00
`XGBoost`	95.67±0.27	92.40±0.48	95.41±0.39	95.88±0.34	2.00
`RF`	95.52±0.36	92.14±0.63	95.28±0.46	95.71±0.40	3.00
`ET`	95.40±0.40	91.94±0.69	95.13±0.43	95.63±0.47	4.00
`GBC`	95.26±0.31	91.66±0.54	94.82±0.41	95.63±0.35	5.00

Notes. Metrics obtained using the default probability threshold of 0.5. Algorithms are sorted by decreasing recall values.
For display purposes, all metrics have been multiplied by 100**.**
Uncertainties show standard deviation of metrics obtained across all 10 **training folds (cf. Sect. 3.3).**

Table 4. Resulting metrics of AGN-galaxy classification model for the test subset and the labelled sources in S82 using two different threshold values, as described in Sect. 4.1. HETDEX and **S82** pipeline results are described in Sect. 4.4.

Subset	Threshold	F_β (×100)	MCC (×100)	Precision (×100)	Recall (×100)
HETDEX-test	Naive	95.37±0.36	91.81±0.67	97.47±0.69	95.89±2.27
	PR	95.42±0.38	91.85±0.70	94.49±0.65	96.21±0.43
S82-label	Naive	94.15±0.44	70.54±2.02	95.16±0.41	93.33±0.66
	PR	94.37±0.36	70.67±1.72	94.81±0.40	94.01±0.59
HETDEX-pipe	Naive	95.37±0.36	91.81±0.67	97.47±0.69	95.89±2.27
	PR	95.42±0.38	91.85±0.70	94.49±0.65	96.21±0.43
S82-pipe	Naive	94.15±0.44	70.54±2.02	95.16±0.41	93.33±0.66
	PR	94.37±0.36	70.67±1.72	94.81±0.40	94.01±0.59

Notes. All metrics have been multiplied by 100**.**
Uncertainties show standard deviation of metrics obtained across all 10 **training folds (cf. Sect. 3.3).**

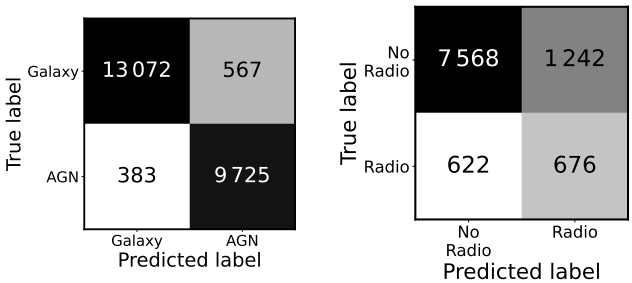
Overall, the model is able to separate AGN from galaxies with a very high (recall $\gtrsim$94%) success rate. **A comparison with traditional colour-colour criteria for AGN selection is presented in Sect. 5.1.1. In particular, Table 15 displays metrics for such criteria. Our classification model can recover, in the HETDEX field,** 15% **and** 59% **more AGN than said formulae. In the S82 field, these differences range between** 17% **and** 61%**. Such differences highlight the fact that most of the information that separates AGN from galaxies is traced by the selected features (mostly colours). Also, the increase in the recovery rate underlines the importance of using photometric information from several bands for such task, as opposed to traditional colour-colour criteria.**

4.2. Radio detection

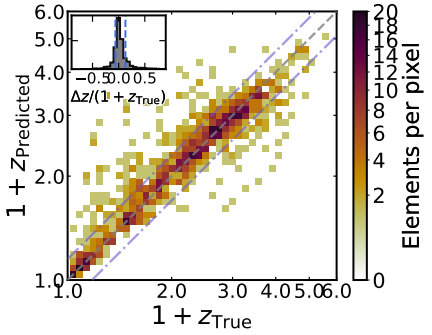
Training of the radio detection model was applied only to sources confirmed to be AGN (`class == 1`). Feature selection was applied to the train+validation subset, with 36 387 confirmed AGN. The target feature is `LOFAR_detect` and the base of selected features are: `band_num`, `W4mag`, `g_r`, `g_i`, `r_i`, `r_z`, `i_z`, `z_y`, `z_W1`, `y_J`, `y_W1`, `J_H`, `H_K`, `K_W3`, `K_W4`, `W1_W2`, and `W2_W3`.

The performance of the tested algorithms is shown in Table 5. In this case, `GBC` shows the highest mean rank. For this reason, we used it as the meta-learner and `XGBoost`, `CatBoost`, `RF`, and `ET` were selected as base-learners.

The optimal threshold for this model is found to be ~0.20460.



(a) AGN-galaxy classification (b) Radio detection on AGN



(c) Redshift on radio AGN

Fig. 6. Performance of individual models (AGN-Galaxy classification, radio-detectability classification and redshift regression) when applied to the HETDEX test subset. (a): confusion matrix for AGN-galaxy classification. (b): Same as (a), but for radio detection. (c): Density plot comparison between original and the predicted redshifts. Grey, dashed line shows the 1:1 relation while dot-dashed lines show the limits for outliers (cf. Eq. 18). Inset displays the distribution of Δz^{N} with a $<\Delta z^{N}>$ = 0.0442.

Table 5. Best performing models the radio detection classification

Model	F_β (×100)	MCC (×100)	Precision (×100)	Recall (×100)	Rank
XGBoost	29.98±2.29	29.81±2.17	56.74±2.93	21.61±2.00	2.75
CatBoost	29.57±1.62	30.56±1.71	60.10±2.85	20.85±1.36	2.25
GBC	29.60±1.66	31.31±1.93	62.55±3.95	20.66±1.40	1.75
RF	29.16±2.47	30.26±2.65	60.03±3.73	20.48±1.96	3.75
ET	28.40±1.27	29.73±1.47	60.06±2.85	19.80±1.05	4.50

Notes. Values and uncertainties as in Table 3.

Finally, the stacked model metrics and confusion matrix are shown in Table 6, for PR-optimised and naive thresholds, and in Fig. 6 respectively.

Table 6. Resulting metrics of the radio detection model on the test subset and the labelled sources in S82 using two different threshold values, as explained in Sect. 4.2. HETDEX and **S82** pipeline results shown as part of the discussion in Sect. 4.4.

Subset	Threshold	F_β (×100)	MCC (×100)	Precision (×100)	Recall (×100)
HETDEX-test	Naive	24.87±2.94	27.36±3.46	60.61±8.18	16.72±2.31
	PR	42.88±2.93	32.47±3.49	35.28±2.74	52.16±3.59
S82-label	Naive	27.15±2.28	23.36±2.27	25.72±1.91	28.47±3.24
	PR	21.62±1.20	19.37±1.64	12.29±0.73	58.16±3.06
HETDEX-pipe	Naive	24.37±3.53	26.93±4.18	59.36±7.17	16.38±2.63
	PR	41.57±4.16	31.67±4.81	34.65±3.24	49.80±5.85
S82-pipe	Naive	26.52±5.44	23.29±5.73	25.71±5.89	27.72±5.21
	PR	20.19±2.84	18.40±4.07	11.45±1.58	54.78±8.44

Notes. Values and uncertainties as in Table 4.

Table 7. Results of initial fit for redshift value prediction

Model	σ_{MAD} (×100)	σ_{NMAD} (×100)	σ_z (×100)	σ_z^N (×100)	η (×100)	Rank
RF	17.88±1.41	07.95±0.50	42.02±5.28	19.38±2.44	19.51±1.98	2.0
ET	18.53±1.03	08.42±0.43	41.12±4.16	18.65±2.26	19.24±1.16	1.8
CatBoost	21.71±1.38	10.08±0.47	40.35±3.03	18.52±1.39	21.93±1.55	2.2
XGBoost	22.89±1.05	10.84±0.78	43.14±3.99	19.62±1.78	24.15±1.84	4.0
GBR	27.73±1.57	12.72±0.74	44.82±3.80	20.41±1.67	28.67±2.25	5.0

Notes. Algorithms sorted by increasing σ_{MAD} values.
Uncertainties as in Table 3.

Table 8. Redshift prediction metrics for the test subset from HETDEX and **S82** labelled sources as discussed in Sect. 4.4.

Subset	σ_{MAD} (×100)	σ_{NMAD} (×100)	σ_z (×100)	σ_z^N (×100)	η (×100)
HETDEX-test	16.54±2.55	7.27±0.99	41.14±09.97	20.56±5.98	19.03±3.35
S82-label	18.66±2.26	9.28±1.37	51.08±11.62	24.69±4.36	24.29±4.68
HETDEX-pipe-Naive	08.11±3.95	5.42±2.19	32.00±12.27	20.97±9.69	19.01±8.22
HETDEX-pipe-PR	15.86±1.77	7.17±0.81	37.80±03.06	22.93±2.73	18.91±1.59
S82-pipe-Naive	15.17±2.70	9.14±1.23	43.05±07.20	24.32±5.00	24.09±4.52
S82-pipe-PR	20.71±1.23	9.84±0.56	45.14±04.42	26.14±3.77	25.18±2.26

Notes. Values and uncertainties as in Table 4.

4.3. Redshift predictions

The redshift value prediction model was applied to sources confirmed to be radio-detected AGN (i.e. `class == 1` and `radio_detect == 1`). Feature selection (cf. Sect. 3.1) was applied to the train+validation subset, with 4 612 sources, leading to the selection of 17 features. The target feature is `Z` and the selected base features are `band_num`, `W4mag`, `g_r`, `g_W3`, `r_i`, `r_z`, `i_z`, `i_y`, `z_y`, `y_J`, `y_W1`, `J_H`, `H_K`, `K_W3`, `K_W4`, `W1_W2`, and `W2_W3`.

For the redshift prediction, the tested algorithms performed as shown in Table 7. Based on their mean rank values, `RF`, `CatBoost`, `XGBoost`, and `GBR` were selected as base learners and `ET` (which shows the best σ_{MAD} value of the two models with the best rank) was used as meta-learner. The redshift regression metrics of the stacked model are presented in Table 8. Likewise, the comparison between the original and predicted redshifts is shown in the lower panel of Fig. 6.

4.4. Prediction pipeline

The sequential combination of the models described in Sect. 3 defines the pipeline for the prediction of radio-detectable AGN and their redshift. As separate tasks, the pipeline was applied to the labelled sources in the HETDEX testing subset, to the labelled sources in **S82**, and to all unlabelled sources across both fields. **S82** provides an independent test of the pipeline as no data in this field was used for training the different models. A full candidate catalogue is extracted from this exercise and based on the unlabelled datasets.

As the metrics discussed in the previous sections correspond to each individual model, new –combined– metrics, based on the knowledge for labelled sources, are calculated for HETDEX and **S82** and presented in Fig. 8 and Tables 8 and 9. Overall, we observe worse combined metrics with respect to the ones calculated for individual models (**e.g. recall of** 45% **for HETDEX and** 47% **for S82**). This degradation might be understood by the fact that the pipeline is composed of three sequential models. Each additional step is fed with sources classified by the previous algorithm. And some of these sources might not be similar,

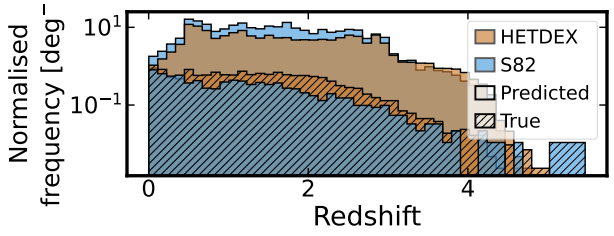


Fig. 7. Redshift density distribution of the predicted radio-AGN within the unlabelled sources (clean histograms) in HETDEX (ochre histograms) and **S82** (blue histograms) and true redshifts from labelled radio-AGN (dashed histograms).

Table 9. Results of application of radio AGN prediction pipeline to the labelled sources in the HETDEX and **S82** fields.

Subset	Threshold	F_β (×100)	MCC (×100)	Precision (×100)	Recall (×100)
HETDEX-test	Naive	20.68±3.17	24.93±3.72	52.34±6.56	13.79±2.27
	PR	37.99±2.59	33.66±2.79	32.20±2.72	44.61±2.46
S82-label	Naive	24.08±3.44	21.43±3.53	25.44±3.64	23.07±3.72
	PR	19.42±2.31	17.23±3.08	11.33±1.32	47.36±6.22

Notes. Values and uncertainties as in Table 3.

in terms of features, to those used for training, thus adding noise to the output of such model. A small sample of the output of the pipeline for five high-z labelled radio AGN sources in HETDEX and **S82** are shown in Tables 10 and 11 respectively.

The application of the prediction pipeline to the unlabelled sources from the HETDEX field led to 9 974 990 predicted AGN, from which 68 252 were predicted to be radio detectable. The pipeline predicts, as well, 2 073 997 AGN in the unlabelled data from **S82**, being 22 445 of them candidates to be detected in the radio (to the detection level of LoTSS). The distribution of the predicted redshifts for radio-AGN in HETDEX and **S82** is presented in Fig. 7. The pipeline outputs for a small sample of the predicted radio AGN are presented in Tables 12 and 13 for HETDEX and **S82** respectively.

Section 5 explores the comparison of these results with previous works in the literature and discusses the main drivers (i.e. features) for the detection of these radio-AGN.

4.5. No-skill classification

As presented in Sect. 3.2.1, Eqs. 6–9 show the base results for a classification with no skill. Table 14 presents the scores generated by using this technique. These values are the base from which any improvement can be assessed.

Subsets and prediction modes displayed in Table 14 coincide with those exhibited in Tables 4, 6, and 9. For instance, in the test HETDEX sub-sample, ∼43% of sources are labelled as AGN. From all AGN, ∼13% of them have radio detections. This can be summarised stating that ∼6% of all sources in the test sub-sample are radio-detected AGN.

5. Discussion

5.1. Comparison with previous prediction or detection works

In this subsection, we provide a few examples of related published works as well as plausible explanations for observed discrepancies when these are present. This comparison attempts to be representative of the literature on the subject but does not intends to be complete in any way.

5.1.1. AGN detection prediction

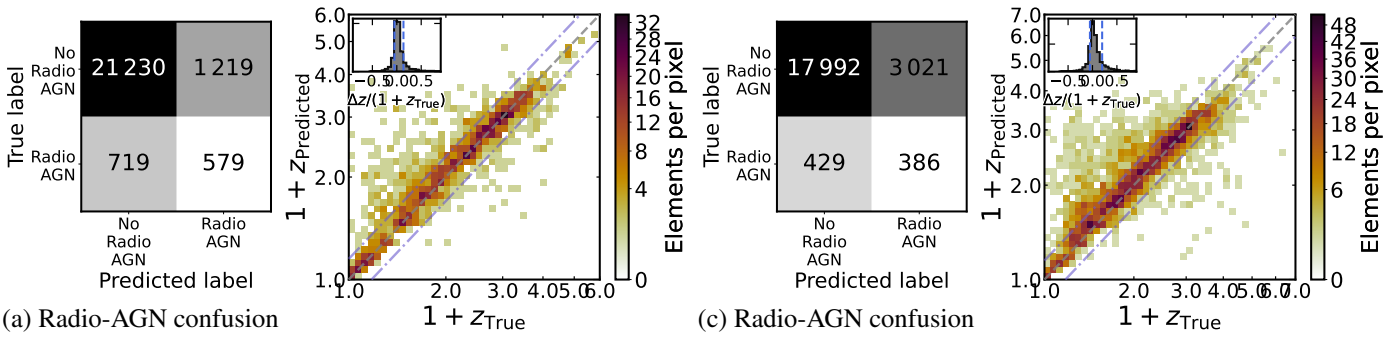
We separate the comparison with previously published results between traditional and ML methodologies in order to understand the significance of our results and ways for future improvement.

Traditional AGN selection methods are based on the comparison of the measured Spectral Energy Distribution (SED) photometry to a template library (Walcher et al. 2011). A recent example of its application is presented by Thorne et al. (2022) where best fit classifications were calculated for more than 700 000 galaxies in the D10 field of the Deep Extragalactic VIsible Legacy Survey (DEVILS; Davies et al. 2018) and the Galaxy and Mass Assembly survey (GAMA; Driver et al. 2011; Liske et al. 2015). The 91% recovery rate of AGN, selected through various means (X-ray measurements, narrow and broad emission lines, and mid-infrared colours), is very much in line with our findings in **S82**, where our rate (recall) reaches 89%.

Traditional methods also encompass the colour-based selection of AGN. While less precise, they provide access to a much larger base of candidates with a very low computational cost. We implemented some of the most common colour criteria on the data from **S82**. Of particular interest is the predicting power of the mid-IR colour selection due to its potential to detect hidden or heavily obscured AGN activity. Based on WISE (Wright et al. 2010) data, Stern et al. (2012, S12) proposed a threshold at W1 - W2 $\geq$ 0.8 to separate AGN from non-AGN using data from AGN in the COSMOS field (Scoville et al. 2007). A more stringent criterion was developed by Mateos et al. (2012, M12), the AGN wedge, which can be defined by the sources located inside the region defined by the relations W1 - W2 $<$ 0.315 $\times$ (W2 - W3) + 0.791, W1 - W2 $>$ 0.315 $\times$ (W2 - W3) – 0.222, and W1 - W2 $>$ −3.172 $\times$ (W2 - W3) + 7.624. In order to define this wedge, they used data from X-ray selected AGN over an area of 44.43 deg^2 in the northern sky. Mingo et al. (2016, M16) cross-correlated data from WISE observations with X-ray and radio surveys creating a sample of star-forming galaxies and AGN in the northern sky. They developed individual relations to separate classes of galaxies and AGN in the W1 - W2, W2 - W3 space and, for AGN the criterion, the relation is W1 - W2 $\geq$ 0.5 and W2 - W3 $<$ 4.4. More recently, Blecha et al. (2018, B18) analysed the quality of mid-IR colour selection methods for the identification of obscured AGN involved in mergers. Using hydrodynamic simulations for the evolution of AGN in galaxy mergers, they developed a selection criterion from WISE colours which is shown to be able to separate, with high reliability, starburst galaxies from AGN. The expressions have the form W1 - W2 $>$ 0.5, W2 - W3 $>$ 2.2, and W1 - W2 $>$ 2 $\times$ (W2 - W3) – 8.9.

The results from the application of these criteria to our samples in the testing subset and in the labelled sources of **S82** field are summarised in Table 15 and a graphical representation of the boundaries they create in their respective parameter spaces is presented in Fig. 9.

Table 15 shows that previous colour-colour criteria have been designed and calibrated to have very high precision values. Most of the sources deemed to be AGN by them are, indeed, of such class. Despite being tuned to maximise their recall (and F_β to a lesser extent), our classifier, and the criterion derived from it, still show precision values compatible with those of such criteria. This result underlines the power of ML methods. They can be on a par with traditional colour-colour criteria and excel in additional metrics.



(a) Radio-AGN confusion matrix

(b) Predicted-True z comparison

(c) Radio-AGN confusion matrix

(d) Predicted-True z comparison

Fig. 8. Combined confusion matrices and True/predicted redshift density plot for the full radio AGN detection prediction computed using the testing sub-set from HETDEX (panels (a) and (b)) and the known labelled sources from **S82** (panels (c) and (d)).

Table 10. Predicted and original properties for the 5 sources in testing subset with the highest redshift predicted Radio AGN. Sources are sorted by decreasing predicted redshift. A description of the columns is presented in Appendix G.

ID	RA_ICRS (deg)	DE_ICRS (deg)	band_num	class	Score_AGN	Prob_AGN	LOFAR_detect	Score_radio	Prob_radio	Score_rAGN	Prob_rAGN	z	pred_z
9898717	203.016113	55.518097	9	1.0	0.500082	0.954114	0	0.390861	0.375122	0.195462	0.357909	4.738	4.3679
168686	164.769135	45.806320	8	1.0	0.500048	0.858157	0	0.450279	0.418719	0.225161	0.359326	4.893	4.1733
14437074	213.226517	54.236343	9	1.0	0.500090	0.965187	0	0.251632	0.263746	0.125839	0.254564	4.326	4.0475
10408176	188.163651	52.880898	9	1.0	0.500012	0.622448	0	0.604838	0.526003	0.302426	0.327410	4.340	3.9553
12612753	227.216370	51.941029	9	1.0	0.500055	0.887909	0	0.364423	0.355080	0.182231	0.315278	3.795	3.8797

Table 11. Predicted and original properties for the 5 sources in **S82** with the highest predicted redshift on the labelled sources predicted to be Radio AGN. Sources are sorted by decreasing predicted redshift. A description of the columns is presented in Appendix G.

ID	RA_ICRS (deg)	DE_ICRS (deg)	band_num	class	Score_AGN	Prob_AGN	radio_detect	Score_radio	Prob_radio	Score_rAGN	Prob_rAGN	z	pred_z
1406323	32.679794	-0.305035	6	1.0	0.500050	0.866373	1	0.185842	0.204867	0.092930	0.177491	4.650	4.4986
326139	33.580879	-1.121398	8	1.0	0.500040	0.822622	0	0.208769	0.225946	0.104393	0.185868	4.600	4.3785
633752	12.526446	-0.888660	9	1.0	0.500035	0.793882	0	0.206182	0.223600	0.103098	0.177512	4.310	4.2946
2834844	344.101440	0.789000	7	1.0	0.500062	0.909395	0	0.375735	0.363709	0.187891	0.330756	4.099	4.0635
3191865	31.881712	1.063655	9	1.0	0.500087	0.962260	0	0.264210	0.274477	0.132128	0.264118	3.841	4.0509

Figure 9 is constructed as **a confusion matrix**, plotting in each quadrant the whole WISE population in the background and in colour **contours** the corresponding fraction of the testing set (TP, TN, FP, and FN**, see Fig. 6a and Sect. 3.2.1**). As expected, our pipeline is able to separate with high confidence sources which are closer to the AGN or the galaxy locus (TP and TN) while **sources in the FN and FP quadrant show a different situation. AGN predicted to be galaxies (FN,** 1.6% **of sources for HETDEX, and** 4.9% **for S82) are located in the galaxy region of the colour-colour diagram. On the opposite corner of the plot, galaxies predicted to be AGN (FP,** 2.4% **of sources for HETDEX, and** 4.2% **for S82) cover the areas of AGN and galaxies uniformly. FN sources might be sources that are identified as AGN by means not included in our feature set (e.g. X-ray, radio emission). FP sources, alternatively, might be galaxies with extreme properties, similar to AGN.**

For the case of ML-based models for AGN-galaxy classification, several analyses have been published in recent years. An example of their application is provided in Clarke et al. (2020) where a Random Forest model for the classification of stars, galaxies and AGN using photometric data was trained from more than 3 000 000 sources in the SDSS (DR15; Aguado et al. 2019) and WISE with associated spectroscopic observations. Close to 400 000 sources have a quasar spectroscopic label and from the application of their model to a validation subset, they obtain a recall of 0.929 and F1-score of 0.943 for the quasar classification. These scores are of the same order as the ones obtained when applying our AGN-Galaxy model to the testing set (see Table 4). Thus, and despite using an order of magnitude fewer sources for the full training and validation process, our model can achieve equivalently good scores.

Expanding on Clarke et al. (2020), Cunha & Humphrey (2022) built a ML pipeline, `SHEEP`, for the classification of sources into stars, galaxies and QSO. In contrast to Clarke et al. (2020) or the pipeline described here, the first step in their analysis is the redshift prediction, which is used as part of the training features by the subsequent classifiers. They extracted WISE and SDSS (DR15; Aguado et al. 2019) photometric data for almost 3 500 000 sources classified as stars, galaxies or QSO. The application of their pipeline to sources predicted to be QSO led to a recall of 0.960 and an F1 score of 0.967. The improved scores in their pipeline might be a consequence not only of the slightly larger pool of sources, but also the inclusion of the coordinates of the sources (RA, Dec) and the predicted redshift values as features in the training.

A test with a larger number of ML methods was performed by Poliszczuk et al. (2021). For training, they used optical and infrared data from close to 1 500 sources (galaxies and AGN) located at the AKARI North Ecliptic Pole (NEP) Wide-field (Lee et al. 2009; Kim et al. 2012) covering a 5.4 deg^2 area. They tested `LR`, `SVM`, `RF`, `ET`, and `XGBoost` including the possibility of generalised stacking. In general, they obtained results with F1-scores between 0.60 – 0.70 and recall values in the range of 50% – 80%. These values, lower than the works described here, can be fully understood given the small size of the training sample. A larger photometric sample covers a wider range of the parameter space which significantly helps the metrics of any given model.

Table 12. Predicted and original properties for the 5 sources in the HETDEX field with the highest predicted redshift on the unlabelled sources predicted to be Radio AGN. A description of the columns is presented in Appendix G.

ID	RA_ICRS (deg)	DE_ICRS (deg)	band_num	Score_AGN	Prob_AGN	radio_detect	Score_radio	Prob_radio	Score_rAGN	Prob_rAGN	pred_z
9544254	201.309235	53.746429	6	0.500007	0.578804	0	0.351672	0.345250	0.175838	0.199832	4.7114
12355845	220.838120	50.319016	5	0.500007	0.578804	0	0.937123	0.794128	0.468568	0.459644	4.6064
13814216	219.839142	52.660328	7	0.500015	0.650248	0	0.213846	0.230529	0.106926	0.149901	4.5622
6698239	184.694901	49.063766	5	0.499995	0.467527	0	0.799085	0.662753	0.399538	0.309855	4.5483
2951011	175.882446	55.497799	5	0.500008	0.589419	0	0.823295	0.681768	0.411654	0.401847	4.5320

Table 13. Predicted and original properties for the 5 sources in **S82** with the highest predicted redshift on the unlabelled sources predicted to be Radio AGN. A description of the columns is presented in Appendix G.

ID	RA_ICRS (deg)	DE_ICRS (deg)	band_num	Score_AGN	Prob_AGN	radio_detect	Score_radio	Prob_radio	Score_rAGN	Prob_rAGN	pred_z
3244450	26.276423	1.104065	7	0.500002	0.531172	0	0.542061	0.483128	0.271031	0.256624	4.3938
1062270	11.744675	-0.562642	7	0.499982	0.356043	0	0.196326	0.214586	0.098159	0.076402	4.3563
3261269	28.882526	1.117103	7	0.500011	0.608660	0	0.354936	0.347777	0.177472	0.211678	4.3153
1466227	18.157259	-0.258997	5	0.500013	0.630968	0	0.456207	0.422973	0.228110	0.266882	4.3146
1134866	11.304936	-0.507943	7	0.500011	0.616439	0	0.226178	0.241539	0.113091	0.148894	4.3140

Table 14. Results of no-skill selection of sources in different stages of pipeline to the labelled sources in the HETDEX test subset and **S82** fields.

Subset	Prediction	F_β (×100)	MCC (×100)	Precision (×100)	Recall (×100)
HETDEX	AGN-galaxy	42.57	0.00	42.57	42.57
	Radio-detection (label)	12.84	0.00	12.84	12.84
	Radio AGN	05.47	0.00	05.47	05.47
S82	AGN-galaxy	81.29	0.00	81.29	81.29
	Radio-detection (label)	04.59	0.00	04.59	04.59
	Radio AGN	03.73	0.00	03.73	03.73

Notes. All metrics have been multiplied by 100**.**

Table 15. Results of application of several AGN detection criteria to our testing subset and the labelled sources from the **S82** field.

HETDEX test set				
Method[a]	F_β (×100)	MCC (×100)	Precision (×100)	Recall (×100)
S12	86.10	78.78	93.98	80.51
M12	51.80	49.71	98.87	37.18
M16	67.21	61.30	97.48	53.48
B18	82.14	75.76	97.54	72.66
This work	92.71	87.64	94.00	91.67

S82 (labelled)				
Method	F_β (×100)	MCC (×100)	Precision (×100)	Recall (×100)
S12	83.59	45.47	93.93	76.62
M12	46.80	28.22	99.59	32.54
M16	64.69	37.76	98.80	50.32
B18	79.71	51.07	98.72	68.77
This work	90.63	58.53	94.15	87.91

Notes.[a] Naming codes for the used methods are described in the main text (cf. Sect. 5.1.1). Last row of each sub-table corresponds to the criterion derived in this work (as described in Sect 5.3.1). **All metrics have been multiplied by** 100**.**

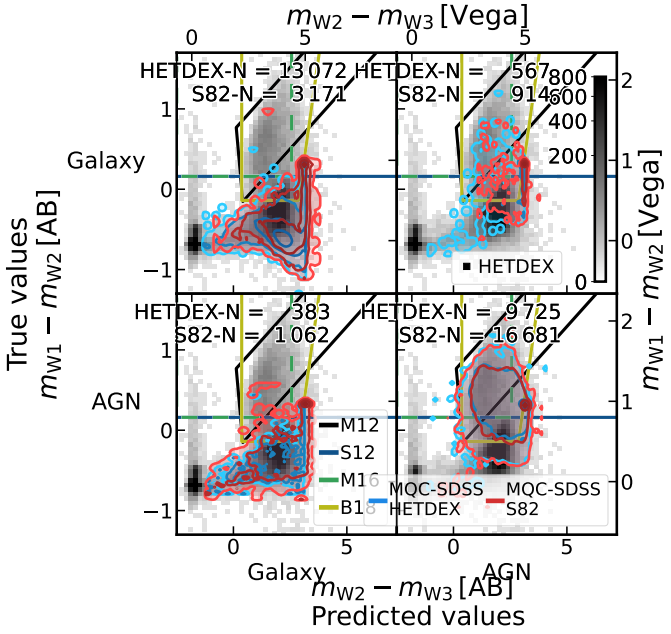


Fig. 9. W1 - W2, W2 - W3 colour-colour diagrams for sources in the testing subset, from HETDEX, and labelled sources from **S82** given their position in the AGN-galaxy confusion matrix (see, for HETDEX, rightmost panel of Fig. 8). In the background, density plot of all CW-detected sources in the full HETDEX field sample is displayed. Colour of each square represents the number of sources in that position of parameter space, with darker squares having more sources (as defined in the colorbar of the upper-right panel). Contours represent distribution of sources for each of the aforementioned subsets at 1, 2, 3, and **4** σ levels (shades of blue, for testing set and shades of red for labelled **S82** sources). Coloured, solid lines display limits from the criteria for the detection of AGN described in Sect. 5.1.1.

5.1.2. Radio detection prediction

We have not found in the literature any work attempting the prediction of AGN radio detection at any level and therefore this is the first attempt at doing so. In the literature we do find several correlations between the AGN radio emission (flux) and that at other wavelengths (e.g. with infrared emission, Helou et al. 1985; Condon 1992) and substantial effort has been done towards classifying radio galaxies based upon their morphology (e.g. Aniyan & Thorat 2017; Wu et al. 2019, FRI, FRII, bent jets, etc.) and its connection to environment (Miley & De Breuck 2008; Magliocchetti 2022). None of these extensive works has directly focused on the a priori presence or absence of radio emission above a certain threshold. Therefore, the results presented here are the first attempt at such an effort.

The ~2x success rate of the pipeline to identify radio emission in AGN (~44.61% recall and ~32.20% precision; see Table 9) with the respect to a 'no-skill' or random (≲30%) se-

lection, provides the opportunity to understand what the model has learned from the data and, therefore, gain some insight into the nature or triggering mechanisms of the radio emission. We, therefore, reserve the discussion of the most important features, and the linked physical processes, driving the pipeline improved predictions to Sect. 5.3.1.

5.1.3. Redshift value prediction

We compare our results to that of Ananna et al. (2017, Stripe 82X) where the authors analysed multi-wavelength data from more than 6 100 X-ray detected AGN from the 31.3 deg^2 of the Stripe 82X survey. They obtained photometric redshifts for almost 6 000 of these sources using the template-based fitting code `LePhare` (Arnouts et al. 1999; Ilbert et al. 2006). Their results present a normalised median absolute deviation of $\sigma_{\rm NMAD}$=0.062 and an outlier fraction of η=13.69%, values which are similar to our results in HETDEX and **S82** except for a better outlier fraction (**as shown in Table 8,** we obtain η_{S82} = 25.18%, $\sigma_{\rm NMAD}^{\rm HETDEX}$=0.071, and $\eta^{\rm HETDEX}$=18.9%).

On the ML side, we compare our results to those produced by Carvajal et al. (2021) **in S82**, with $\sigma_{\rm NMAD}$ = 0.1197 and η = 29.72%, and find that our redshift prediction model improves by at least 25% for any given metric. The source of improvement is probably **many-fold. First, it might be** related to the different sets of features used (colours vs ratios) and**, second,** the more specific population of radio-AGN used to train our models. **Carvajal et al. (2021) used a limited set of colours to train their model, while we have allowed the use of all available combinations of magnitudes (Sect. 2.2). Additionally, their redshift model was trained on all available AGN in HETDEX, while we have trained (and tested) it only with radio-detected AGN. Using a more constrained sample reduces the likelihood of handling sources that are too different in the parameter space.**

Another example of the use of ML for AGN redshift prediction has been presented by Luken et al. (2019). They studied the use of the k-nearest neighbours algorithm `KNN` (Cover & Hart 1967), a non-parametric supervised learning approach, to derive redshift values for radio-detectable sources. They combined 1.4 GHz radio measurements, infrared, and optical photometry in the European Large Area ISO Survey-South 1 (ELAIS-S1; Oliver et al. 2000) and extended Chandra Deep Field South (eCDFS; Lehmer et al. 2005) fields, matching their sensitivities and depths to the expected values in the Evolutionary Map of the Universe (EMU; Norris et al. 2011). From the different experiments they run, their resulting NMAD values are in the range $\sigma_{\rm NMAD} = 0.05 - 0.06$, and their outlier fraction can be found between η = 7.35% and η = 13.88%. As an extension to the previous results, Luken et al. (2022) analysed multi-wavelength data from radio-detected sources the eCDFS and the ELAIS-S1 fields. Using `KNN` and `RF` methods to predict the redshifts of more than 1 300 RGs, they have developed regression methods that show NMAD values between $\sigma_{\rm NMAD}$ = 0.03 and $\sigma_{\rm NMAD}$ = 0.06, $\sigma_z = 0.10 - 0.19$, and outlier fractions of η = 6.36% and η = 12.75%.

In addition to the previous work, Norris et al. (2019) compared a number of methodologies, mostly related with ML but also `LePhare`, for predicting redshift values for radio sources. They have used more than 45 photometric measurements (including 1.4 GHz fluxes) from different surveys in the COSMOS field. From several settings of features, sensitivities, and parameters, they retrieved redshift predictions with NMAD values between $\sigma_{\rm NMAD}$ = 0.054 and $\sigma_{\rm NMAD}$ = 0.48 and outlier fractions that range between η = 7% and η = 80%. The broad span of obtained values might be due to the combinations of properties for each individual training set (including the use of radio or X-ray measurements, the selection depth, and others) and to the size of these sets, which was small for ML purposes (less than 400 sources). The slightly better results can be understood given the heavily populated photometric data available in COSMOS.

Specifically related to HETDEX, it is possible to compare our results to those from Duncan et al. (2019). They use a hybrid photometric redshift approach combining traditional template fitting redshift determination and ML-based methods. In particular, they implemented a Gaussian Process (GP) algorithm, which is able to model both the intrinsic noise and the uncertainties of the training features. Their redshift prediction analysis of AGN sources with a spectroscopic redshift detected in the LoTSS DR1 (6, 811 sources) found a NMAD value of $\sigma_{\rm NMAD}$ = 0.102 and an outlier fraction of η = 26.6%. The differences between these results and those obtained from the application of our models (individually and as part of the prediction pipeline) might be due to the differences in the creation of the training sets. Duncan et al. (2019) use information from all available sources in the HETDEX field for training the redshift GP whilst our redshift model has been only trained on radio-detected AGN, giving it the opportunity to focus its parameter exploration only on these sources.

Finally, Cunha & Humphrey (2022) also produced photometric redshift predictions for almost 3 500 000 sources (stars, galaxies, and QSO) as part of their pipeline (see Sect. 5.1.1). They combined three algorithms for their predictions: `XGBoost`, `CatBoost`, and `LightGBM` (Ke et al. 2017). This procedure leads to $\sigma_{\rm NMAD}$ = 0.018 and η = 2%. As with previous examples, the differences with our results can be a consequence of the number of training samples. Also, in the case of Cunha & Humphrey (2022), they applied an additional post-processing step to the redshift predictions attempting to predict and understand the appearance of catastrophic outliers.

5.2. Influence of data imputation

One effect which might influence the training of the models and, consequently, the prediction for new sources is related to the imputation of missing values (cf. Sect. 2.1). In Fig. 10, we have plotted the distributions of predicted scores (for classification models) and predicted redshift values as a function of the number of measured bands (`band_num`) for each step of the pipeline as applied to **sources predicted to be of each class in** the test sub-set.

The top panel of Fig. 10 shows the influence of the degree of imputation in the classification between AGN and galaxies. For most of the bins, **probabilities for predicted galaxies** are distributed close to 0.0, without any noticeable trend. **In the case of predicted AGN, the combination of low number of sources and high degree of imputation** (`band_num` < 5) **lead to low mean probabilities.**

The case of radio detection classification is somewhat different. **Given the number and distribution of sources per bin, it is not possible to extract any strong trend for the probabilities of radio-predicted sources. The absence of evolution with the number of observed bands is stronger for sources predicted to be devoid of radio detection.**

Finally, a stronger effect can be seen with the evolution of predicted redshift values for radio-detectable AGN. Despite the lower number of available sources, it is possible to recognise that sources with higher number of available measurements are

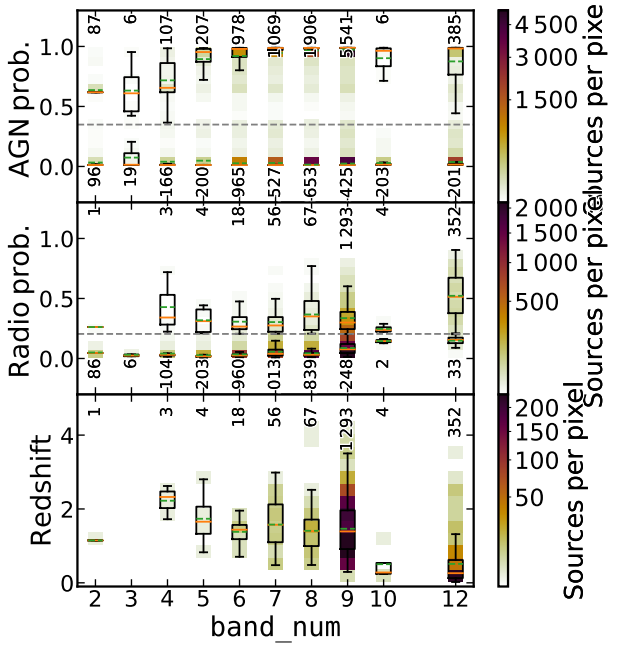


Fig. 10. Evolution of predicted probabilities (top: probability to be AGN, middle: probability of AGN to be detected in radio) and redshift values for radio-detectable AGN (bottom panel) as a function of the number of observed bands **for** sources in test set. **In top panel, sources have been divided between those predicted to be AGN and galaxy. In middle panel, sources are divided between predicted AGN that are predicted to be detected in radio and those predicted to not have radio detection.** Background density plots (following colour coding in colorbars) show location of predicted values. Overlaid boxplots display main statistics for each number of measured bands. Black rectangles encompass sources in second and third quartiles. Vertical lines show the place of sources from first and fourth quartiles. Orange lines represent median value of sample and dashed, green lines indicate their mean values. **Dashed, grey lines show PR thresholds for AGN-galaxy and radio detection classifications. Close to** each boxplot, written values correspond to the number of sources considered to create each set of statistics.

predicted to have lower redshift values. Sources that are closer to us have higher probabilities to be detected in a large number of bands. Thus, it is expected that our model predicts lower redshift values for the most measured sources in the field.

In consequence, Fig. 10 allows us to understand the influence of imputation over the predictions. The most highly affected quantity is the redshift, where large fractions of measured magnitudes are needed to obtain scores that are in line with previous results (cf. Sect. 5.1.3). The AGN-galaxy **and radio detection** classification**s show a mild** influence of imputation in **their** results.

5.3. Model explanations

Given the success of the models and pipeline in classifying AGN, their radio detectability and redshift with the provided set of observables, knowing the relative weights that they have in the decision-making process is of utmost relevance. In this way, physical insight might be gained about the triggers of AGN and radio activity and its connection to their host. Therefore, we have estimated both local and global feature importances for the individual models and the combined pipeline. Global importances were retrieved using the so-called 'decrease in impurity' approach (see, for example, Breiman 2001). Local importances have been determined via Shapley values. A more detailed description of what these importances are and how they are calculated is given in the following sections.

Table 16. Relative importances (rescaled to add to 100) for observed features from the three models combined between meta and base models.

AGN-Galaxy (meta-model: `CatBoost`)					
Feature	Importance	Feature	Importance	Feature	Importance
`W1_W2`	68.945	`H_K`	1.715	`z_W2`	1.026
`W1_W3`	4.753	`y_W1`	1.659	`z_y`	0.722
`g_r`	4.040	`y_W2`	1.513	`W3_W4`	0.669
`r_J`	4.006	`i_y`	1.441	`W4mag`	0.558
`r_i`	3.780	`i_z`	1.366	`H_W3`	0.408
`band_num`	1.842	`y_J`	1.187	`J_H`	0.371
Radio detection (meta-model: `GBC`)					
Feature	Importance	Feature	Importance	Feature	Importance
`W2_W3`	9.609	`y_W1`	7.150	`W4mag`	4.759
`y_J`	8.102	`g_r`	7.123	`K_W4`	2.280
`W1_W2`	8.010	`z_W1`	7.076	`J_H`	1.283
`g_i`	7.446	`r_z`	6.981	`H_K`	1.030
`K_W3`	7.357	`i_z`	6.867	`band_num`	1.018
`z_y`	7.321	`r_i`	6.588		
Redshift prediction (meta-model: `ET`)					
Feature	Importance	Feature	Importance	Feature	Importance
`y_W1`	35.572	`y_J`	3.018	`i_z`	1.215
`W1_W2`	13.526	`r_z`	3.000	`J_H`	1.162
`W2_W3`	12.608	`r_i`	2.896	`g_W3`	1.000
`band_number`	6.358	`z_y`	2.827	`K_W3`	0.925
`H_K`	4.984	`W4mag`	2.784	`K_W4`	0.762
`g_r`	4.954	`i_y`	2.408		

5.3.1. Global feature importances

Overall, mean or global feature importances can be retrieved from models that are based on Decision Trees (e.g. Random Forests and Boosting models, Breiman 2001, 2003). All algorithms selected in this work (`RF`, `CatBoost`, `XGBoost`, `ET`, `GBR`, and `GBC`) belong to these two classes. For each feature, the decrease in impurity (a term frequently used in the literature related to Machine Learning) of the dataset is calculated for all the nodes of the tree in which that feature is used. Features with the highest impurity decrease will be more important for the model (Louppe et al. 2013)[9].

Insight into the decision-making of the pipeline can only rely on the specific weight of the original set of features (see Sect. 3.1). Table 16 presents the ranked combined importances from the observables selected in each of the three sequential models that compose the pipeline. They have been combined using the importances from the meta-learner (as shown in Table 17) and that of base-learners. The derived importances will be dependent on the dataset used, including any imputation for the missing data, and the details of the models, i.e. algorithms used and stacking procedure. We first notice in Table 16 that the order of the features is different for all three models. This difference reinforces the need, as stated in Sect. 3, of developing separate models for each of the prediction stages of this work that would evaluate the best feature weights for the related classification or regression task.

[9] For some models that are not based on Decision Trees, feature importances can be obtained from the coefficients that the training process delivers for each feature. These coefficients are related to the level to which each quantity is scaled to obtain a final prediction (as in the coefficients from a polynomial regression).

Table 17. Relative feature importances (rescaled to add to 100) for base algorithms in each prediction step.

AGN-Galaxy model (CatBoost)			
Feature	Importance	Feature	Importance
gbc	49.709	xgboost	14.046
et	19.403	rf	8.981
Remaining feature importances:			7.861
Radio detection model (GBC)			
Feature	Importance	Feature	Importance
rf	12.024	catboost	7.137
et	7.154	xgboost	6.604
Remaining importances:			67.081
Redshift prediction model (ET)			
Feature	Importance	Feature	Importance
xgboost	25.138	catboost	21.072
gbr	21.864	rf	13.709
Remaining importances:			18.217

For the AGN-galaxy classification model, it is very interesting to note that the most important feature for the predicted probability of a source to be an AGN is the WISE colour W1 - W2 (as well as W1 - W3). This colour is indeed one of the axes of the widely used WISE colour-colour selection, with the second axis being the W2 - W3 colour (cf. Sect 5.1.1). The WISE W3 photometry is though significantly less sensitive than W1, W2 or PS1 (see Fig. 3) and a significant number of sources will be represented as upper limits in such plot (see Table 2). From the importances in Table 16 and the values presented in Fig. 1 we infer that using optical colours could in principle create selection criteria with metrics equivalent to those shown in Table 15 but for a much larger number of sources (100 000 sources for colour plots using W3 vs 4 700 000 sources for colours based in r, i or z magnitudes). We tested this hypothesis and derived a selection criterion in the **g - r** vs W1 - W2 colour-colour plot shown in Fig. 11 using the labelled sources in the test sub-set of the HETDEX field. The results of the application of this criterion to the testing data and to the labelled sources in **S82** is presented in the last row of Table 15. Their limits are defined by the following expressions:

$$\mathbf{g-r} > \mathbf{-0.76}, \quad (21)$$

$$\mathbf{g-r} < \mathbf{1.8}, \quad (22)$$

$$\mathbf{W1-W2} > \mathbf{0.227}\times(\mathbf{g-r})+\mathbf{0.43}, \quad (23)$$

where W1, W2, **g**, and **r** are Vega magnitudes. Our colour criteria provides better and more homogeneous scores across the different metrics with purity (precision) and completeness (recall) above 87%. Avoiding the use of the longer WISE wavelength (W3 and W4), the criteria can be applied to a much larger dataset.

One of the main potential uses of the pipeline is its capability to pinpoint radio-detectable AGN. The global features analysis for the radio detection model shows a high dependence on the near- and mid-IR magnitudes and colours, especially those coming from WISE. As a useful outcome similar to the AGN-Galaxy classification, we can use the most relevant features to build useful plots for the pre-selection of these sources and get insight into the origin of the radio emission. This is the case for the W4 histogram, shown in Fig. 12, where **sources predicted to be** radio-emitting AGN **extend to** brighter measured W4 magnitudes. This added mid-IR flux might be simply due to an increased star formation rates (SFR) in these sources. In fact the 24μm flux is often used, together with that of Hα as a proxy for SFR (Kennicutt et al. 2009). The radio detection for these sources might have a strong component linked to the ongoing SF, especially for the sources with real or predicted redshift below z~1.5. A detailed exploration of the implications that these dependencies might have in our understanding of the triggering of radio emission on AGN, whether related to star formation (SF) or jets, is left for a future publication (Carvajal et al. in preparation).

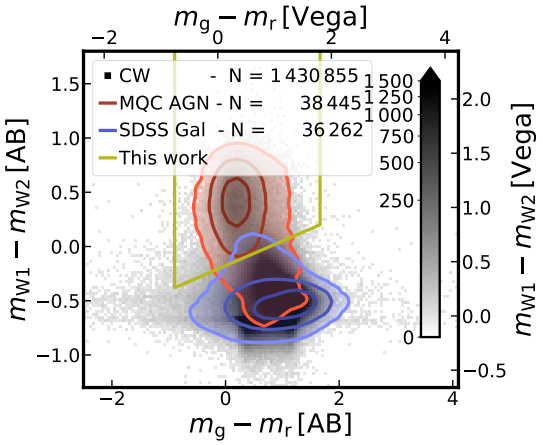


Fig. 11. AGN classification colour-colour plot in the HETDEX field using CW (W1, W2) and PS1 (**g, r**) passbands. Grey-scale density plot include all CW detected and non-imputed sources. Red contours highlight the density distribution of the AGN in the Million QSO catalogue (MQC) and blue contours show the density distribution for the galaxies from SDSS DR16. Contours are located at 1, 2, and 3 σ levels.

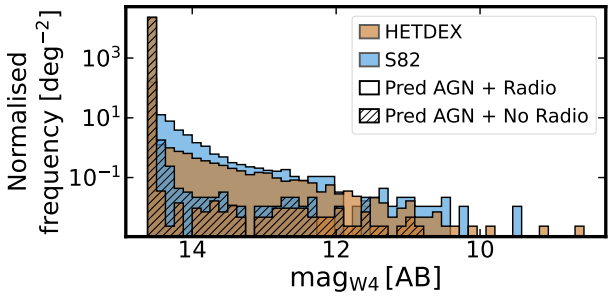


Fig. 12. W4 magnitudes density distribution of the **newly** predicted radio-AGN (clean histograms) in HETDEX (ochre histograms) and **S82** (blue histograms) and W4 magnitudes from predicted AGN that are predicted to not have radio detection (dashed histograms).

Finally, the redshift prediction model shows again that the final estimate is mostly driven by the results of the base learners, accounting for ~82% of the predicting power. The overall combined importance of features shows also in this case a strong dependence on several near-IR colours of which y - W1 and W1 - W2 are the most relevant ones. The model still relies, to a lesser extent, on a broad range of optical features needed to trace the broad range of redshift possibilities ($z \in [0, 6]$).

5.3.2. Local feature importances: Shapley values

As opposed to the global (mean) assessment of feature importances derived from the decrease in impurity, local (i.e. source by source) information on the performance of such features can be obtained from Shapley values. This is a method from coalitional game theory that tells us how to fairly distribute the dividends (the prediction in our case) among the features (Shapley

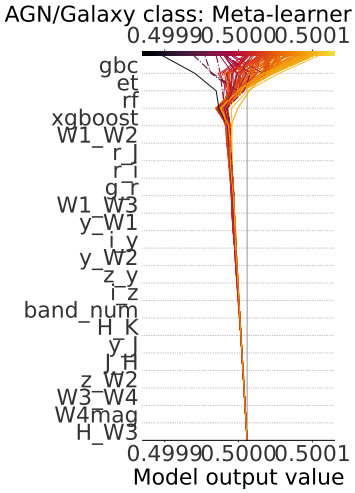


Fig. 13. Decision plot from SHAP values for AGN-Galaxy classification from the 121 high redshift ($z \geq 4$) spectroscopically confirmed AGN in HETDEX. Horizontal axis represents the model's output with a starting value for each source centred on the selected naive threshold for classification. Vertical axis shows features used in the model sorted, from top to bottom, by decreasing mean absolute SHAP value. Each prediction is represented by a coloured line corresponding to its final predicted value as shown by the colorbar at the top. Moving from the bottom of the plot to the top, SHAP values for each feature are added to the previous value in order to highlight how each feature contributes to the overall prediction. Predictions for sources detected by LOFAR are highlighted with a dotted, dashed line.

1953). The previous statement means that the relative influence of each property from the dataset can be derived for individual predictions in the decision made by the model (which is not the same as obtaining causal correlations between features and the target; Ma & Tourani 2020). The combination of Shapley values with several other model explanation methods was used by Lundberg & Lee (2017) to create the SHapley Additive exPlanations (SHAP) values. In this work, SHAP values were calculated using the python package `SHAP`[10] and, in particular, its module for Tree-based predictors (Lundberg et al. 2020). To speed calculations up, the package `FastTreeSHAP`[11] (`v0.1.2`; Yang 2021) was also used, which allows for multi-thread runs.

One way to display these SHAP values is through the so-called decision plots. They can show how individual predictions are driven by the inclusion of each feature. Besides determining the most relevant properties that help the model make a decision, it is possible to detect sources that follow different prediction paths which could be, eventually and upon further examination, labelled as outliers. An example of this decision plot, linked to the AGN-Galaxy classification, is shown in Fig. 13 for a subsample of the high-redshift ($z \geq 4.0$) spectroscopically classified AGN in the HETDEX field (121 sources, regardless of them being part of any sub-set involved in the training or validation of the models). The different features used by the meta-learner are stacked on the vertical axis with increasing weight and these final weight are sumarized in Table 18. Similarly, SHAP decision plots for the radio-detection and redshift prediction are presented in Figs. 14 and 15, respectively.

As it can be seen, for the three models, base learners are amongst the features with the highest influence. This result raises the question of what drives these individual base predictions.

[10] `https://github.com/slundberg/shap`
[11] `https://github.com/linkedin/fasttreeshap`

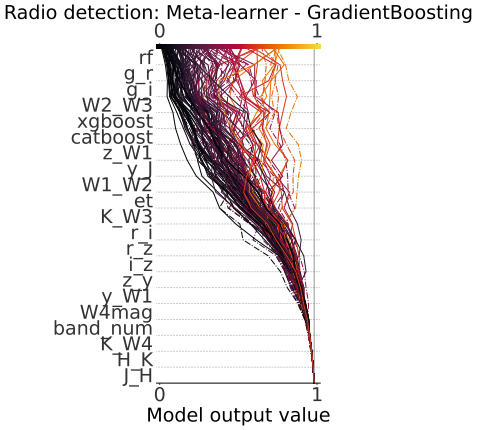


Fig. 14. Decision plot from the SHAP values for all features from the radio detection model in the 121 high redshift ($z \geq 4$) spectroscopically confirmed AGN from HETDEX. Description as in Fig. 13.

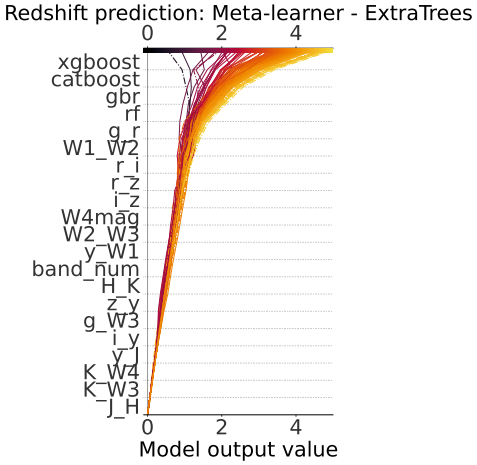


Fig. 15. Decision plot from the SHAP values for all features from the redshift prediction model in the 121 high redshift ($z \geq 4$) spectroscopically confirmed AGN from HETDEX. Description as in Fig. 13.

Appendix F includes SHAP decision plots for all base learners used in this work. Additionally, and to be able to compare these results with the features importances from Sect. 5.3.1, we constructed Table 19, which displays the combined SHAP values of base and meta learners but, in this case, for the same 121 high-redshift confirmed AGN (with 29 of them detected by LoTSS).

Table 18. SHAP values (rescaled to add to 100) for base algorithms in each prediction step for observed features using 121 spectroscopically confirmed AGN at high redshift values ($z > 4$).

AGN-Galaxy model (`CatBoost`)			
Feature	SHAP value	Feature	SHAP value
`gbc`	36.250	`rf`	21.835
`et`	30.034	`xgboost`	7.198
	Remaining SHAP values:		4.683
Radio detection model (`GBC`)			
Feature	SHAP value	Feature	SHAP value
`rf`	11.423	`catboost`	5.696
`xgboost`	7.741	`et`	5.115
	Remaining SHAP values:		70.025
Redshift prediction model (`ET`)			
Feature	SHAP value	Feature	SHAP value
`xgboost`	41.191	`gbr`	13.106
`catboost`	20.297	`rf`	11.648
	Remaining SHAP values:		13.758

Table 19. Combined and normalised (rescaled to add to 100) mean absolute SHAP values for observed features from the three models using 121 spectroscopically confirmed AGN at high redshift values ($z \geq 4$).

AGN-Galaxy model					
Feature	SHAP value	Feature	SHAP value	Feature	SHAP value
W1_W2	32.458	i_y	5.086	z_y	1.591
g_r	11.583	y_W1	4.639	H_W3	1.048
W1_W3	8.816	band_num	4.050	W4mag	0.514
r_i	7.457	y_W2	3.228	H_K	0.466
i_z	6.741	z_W2	2.348	W3_W4	0.466
r_J	6.613	y_J	1.718	J_H	0.178
Radio detection model					
Feature	SHAP value	Feature	SHAP value	Feature	SHAP value
g_i	14.120	z_W1	6.751	W4mag	2.691
W2_W3	13.201	r_i	5.577	band_num	2.661
g_r	12.955	r_z	5.161	K_W4	0.939
y_J	8.224	i_z	4.512	H_K	0.719
K_W3	7.441	z_y	4.121	J_H	0.190
W1_W2	6.874	y_W1	3.864		
Redshift prediction model					
Feature	SHAP value	Feature	SHAP value	Feature	SHAP value
g_r	32.594	z_y	3.557	W4mag	1.639
y_W1	20.770	y_J	3.010	g_W3	1.479
W2_W3	12.462	band_num	2.595	K_W3	0.853
W1_W2	5.692	i_y	2.381	K_W4	0.451
r_i	4.381	H_K	2.230	J_H	0.146
r_z	3.755	i_z	2.005		

Table 19 shows, as Table 16, that the colour W1 - W2 is the most important discriminator between AGN and Galaxies for this specific set of sources. The importance of the rest of the features is mixed: similar colours are located on the top spots (e.g. g - r, W1 - W3 or r - i).

For the radio classification step of the pipeline, we find that features linked to those 121 high-z AGN perform at the same level as for the overall population. The improved metrics with respect to those obtained from the 'no-skill' selection do indicate that the model has learned some connections between the data and the radio emission. Feature importance has changed when compared to the overall population. If the radio emission observed from these sources were exclusively due to SF, this connection would imply SFR of several hundred $M_\odot\,\mathrm{yr}^{-1}$. This explanation can not be completely ruled out from the model side but some contribution of radio emission from the AGN is expected. The detailed analysis of the exact contribution for the SF and AGN component will be left for a forthcoming publication (Carvajal et al. in preparation).

6. Summary and conclusions

With the ultimate intention of better understanding the triggering of radio emission in AGN, in this paper, we have shown that it is possible to build a pipeline to detect AGN, determine their detectability in radio, within a given flux limit, and **predict** their redshift value.

Most importantly, we have described a series of methodologies to understand the driving properties of the different decisions, in particular for the radio detection which is, to our best knowledge, the first attempt at doing so.

We have trained the models using multi-wavelength **photometry** from almost 120 000 spectroscopically identified infrared-detected sources in the HETDEX field and created stacked models **with them**.

These models were applied, sequentially, to 15 018 144 infrared detections in the HETDEX Spring field, arriving to the creation of 68 252 radio AGN candidates with their corresponding predicted redshift values. Additionally, we applied the models to 3 568 478 infrared detections in the **S82** field, obtaining 22 445 new radio AGN candidates with their predicted redshift values.

We have, then, applied a number of analyses on the models to understand the influence of the observed properties over the predictions and their confidence levels. In particular, the use of SHAP values gives the opportunity to extract the influence that the feature set has for each individual prediction.

From the application of the prediction pipeline on labelled and unlabelled sources and the analysis of the predictions and the models themselves, the following conclusions can be drawn.

- Generalised stacking is a useful procedure which collects results from individual ML algorithms into a single model that can outperform each of the individual models**, while preventing** the inclusion of biases **from** individual algorithms. **Proper selection of models and input features, together with detailed probability and threshold calibration maximises the metrics of the final model.**
- Classification between AGN and galaxies derived from our model is in line with previous works. Our pipeline is able to retrieve **a high fraction** of previously-classified AGN from HETDEX (recall = 0.9621, precision = 0.9449) and **from** the **S82** field (recall = 0.9401, precision = 0.9481).
- Radio detection classification for predicted AGN has proven to be highly demanding in terms of data needed for creating the models. Thanks to the use of the techniques shown in this article (i.e. feature creation and selection, generalised stacking, probability calibration, and threshold optimisation), we are able to retrieve previously-known radio-detectable AGN in the HETDEX field (recall = 0.5216, precision = 0.3528) and in the **S82** field (recall = 0.5816, precision = 0.1229). These rates improve **significantly** upon a purely random selection(4 **times better for the HETDEX field and** 13 **times better for S82), showing the power of ML methods for obtaining new RG candidates**.
- The prediction of redshift values for sources classified to be radio-detectable AGN can deliver results that are in line with works that use either traditional or ML methods.
- Our models (classification and regression) can be applied to areas of the sky which have different radio coverage from that used for training without a strong degradation of the prediction results. This feature can lead to the use of our pipeline over very distinct datasets (in radio and multi-wavelength coverage) expecting to recover the sources predicted to be radio-detectable AGN with a high probability.
- Machine Learning models cannot be only used for a direct prediction of a value (or a set of value**s**). They can also be subject to analyses that allow to extract additional results. We took advantage of this fact by using global and local feature importances to derive novel colour-colour AGN selection methods.

With the next generation of observatories already producing source catalogues with an order of magnitude better sensitivity over large areas of the sky than previously (e.g. RACS, EMU, and MIGHTEE; McConnell et al. 2020; Norris et al. 2011; Jarvis et al. 2016, respectively), the need to understand the fraction of those radio detections related to AGN and determine counterparts across wavelengths is more necessary than ever.

Although we developed the pipeline as a tool to better understand the aforementioned issues, we foresee additional possibilities in which the pipeline can be of great use. The first of this

possibilities involves the use of the pipeline to assist with the selection of radio-detectable AGN within any set of observations. This application might turn particularly valuable in recent surveys carried out with MeerKAT (Jonas & MeerKAT Team 2016) or the future SKA where the population at the faintest sources will be dominated by star-forming galaxies. This change needs to use the corresponding data in the training set.

Future developments of the pipeline will concentrate on minimising the existent biases in the training sample as well as in increasing the coverage of the parameter space. We also plan to generalise the pipeline to make it useful for non-radio or galaxy-related research communities. These developments include, for instance, the capability to carry the full analysis for the galactic and stellar populations (i.e. models to determine if a galaxy can be detected in the radio and to predict redshift values for galaxies and non-radio AGN).

In order to **increase** the parameter space of our training sets**, we plan to include** information from radio surveys with different characteristics. Namely, shallower, but with larger area, and less extended but with deeper multi-wavelength data. Similarly, the inclusion of far-IR, X-ray, and multi-survey radio measurements makes part of our efforts to improve detections, not only in radio, but in additional wavelengths.

Acknowledgements. We thank the anonymous referee for their valuable comments and constructive suggestions which have greatly improved the manuscript. The authors would also like to thank insightful comments from P. Papaderos and B. Arsioli. This work was supported by Fundação para a Ciência e a Tecnologia (FCT) through the research grants PTDC/FIS-AST/29245/2017, EXPL/FIS-AST/1085/2021, UID/FIS/04434/2019, UIDB/04434/2020, and UIDP/04434/2020. R.C. acknowledges support from 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). A.H. acknowledges support from contract DL 57/2016/CP1364/CT0002 and an FCT-CAPES funded Transnational Cooperation project "Strategic Partnership in Astrophysics Portugal-Brazil". This publication makes use of data products from the Wide-field Infrared Survey Explorer, which is a joint project of the University of California, Los Angeles, and the Jet Propulsion Laboratory/California Institute of Technology, funded by the National Aeronautics and Space Administration. LOFAR data products were provided by the LOFAR Surveys Key Science project (LSKSP[12]) and were derived from observations with the International LOFAR Telescope (ILT). LOFAR (van Haarlem et al. 2013) is the Low Frequency Array designed and constructed by ASTRON. It has observing, data processing, and data storage facilities in several countries, which are owned by various parties (each with their own funding sources), and which are collectively operated by the ILT foundation under a joint scientific policy. The efforts of the LSKSP have benefited from funding from the European Research Council, NOVA, NWO, CNRS-INSU, the SURF Co-operative, the UK Science and Technology Funding Council and the Jülich Supercomputing Centre. The Pan-STARRS1 Surveys (PS1) and the PS1 public science archive have been made possible through contributions by the Institute for Astronomy, the University of Hawaii, the Pan-STARRS Project Office, the Max-Planck Society and its participating institutes, the Max Planck Institute for Astronomy, Heidelberg and the Max Planck Institute for Extraterrestrial Physics, Garching, The Johns Hopkins University, Durham University, the University of Edinburgh, the Queen's University Belfast, the Harvard-Smithsonian Center for Astrophysics, the Las Cumbres Observatory Global Telescope Network Incorporated, the National Central University of Taiwan, the Space Telescope Science Institute, the National Aeronautics and Space Administration under Grant No. NNX08AR22G issued through the Planetary Science Division of the NASA Science Mission Directorate, the National Science Foundation Grant No. AST-1238877, the University of Maryland, Eötvös Loránd University (ELTE), the Los Alamos National Laboratory, and the Gordon and Betty Moore Foundation. 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 work made use of public data from the Sloan Digital Sky Survey, Data Release 16. Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS-IV acknowledges support and resources from the Center for High Performance Computing at the University of Utah. The SDSS website is `www.sdss.org`. SDSS-IV is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, Center for Astrophysics | Harvard & Smithsonian, the Chilean Participation Group, the French Participation Group, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU) / University of Tokyo, the Korean Participation Group, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatário Nacional / MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt University, and Yale University. This research has made use of NASAs Astrophysics Data System, TOPCAT[13] (Taylor 2005), JupyterLab[14] (Kluyver et al. 2016), "Aladin sky atlas" (`v11.0.24`; Bonnarel et al. 2000) developed at CDS, Strasbourg Observatory, France, and the VizieR catalogue access tool, CDS, Strasbourg, France (DOI : 10.26093/cds/vizier). The original description of the VizieR service was published in Ochsenbein et al. (2000). This work made extensive use of the Python packages `PyCaret`[15] (`v2.3.10`; Ali 2020), `scikit-learn` (`v0.23.2`; Pedregosa et al. 2011), `pandas`[16] (`v1.4.2`; Wes McKinney 2010), `Astropy`[17], a community-developed core Python package for Astronomy (`v5.0`; Astropy Collaboration et al. 2013, 2018, 2022), `Matplotlib` (`v3.5.1`; Hunter 2007), `betacal`[18] (`v1.1.0`), and `CMasher`[19] (`v1.6.3`; van der Velden 2020).

References

Afonso, J., Casanellas, J., Prandoni, I., et al. 2015, in Advancing Astrophysics with the Square Kilometre Array (AASKA14), 71
Aguado, D. S., Ahumada, R., Almeida, A., et al. 2019, ApJS, 240, 23
Ahumada, R., Prieto, C. A., Almeida, A., et al. 2020, ApJS, 249, 3
Alatalo, K., Lacy, M., Lanz, L., et al. 2015, ApJ, 798, 31
Alegre, L., Sabater, J., Best, P., et al. 2022, MNRAS, 516, 4716
Ali, M. 2020, PyCaret: An open source, low-code machine learning library in Python, pyCaret version 2.3
Allen, D. M. 1974, Technometrics, 16, 125
Allison, P. 2001, Missing Data, Quantitative Applications in the Social Sciences (SAGE Publications)
Amarantidis, S., Afonso, J., Messias, H., et al. 2019, MNRAS, 485, 2694
Ananna, T. T., Salvato, M., LaMassa, S., et al. 2017, ApJ, 850, 66
Anbajagane, D., Evrard, A. E., & Farahi, A. 2022, MNRAS, 509, 3441
Aniyan, A. K. & Thorat, K. 2017, ApJS, 230, 20
Annis, J., Soares-Santos, M., Strauss, M. A., et al. 2014, ApJ, 794, 120
Arnouts, S., Cristiani, S., Moscardini, L., et al. 1999, MNRAS, 310, 540
Arsioli, B. & Dedin, P. 2020, MNRAS, 498, 1750
Assef, R. J., Stern, D., Noirot, G., et al. 2018, ApJS, 234, 23
Astropy Collaboration, Price-Whelan, A. M., Lim, P. L., et al. 2022, ApJ, 935, 167
Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, AJ, 156, 123
Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33
Baldwin, J. A., Phillips, M. M., & Terlevich, R. 1981, PASP, 93, 5
Ball, N. M. & Brunner, R. J. 2010, International Journal of Modern Physics D, 19, 1049
Ball, N. M., Brunner, R. J., Myers, A. D., et al. 2008, ApJ, 683, 12
Banfield, J. K., Wong, O. I., Willett, K. W., et al. 2015, MNRAS, 453, 2326
Baron, D. 2019, arXiv e-prints, arXiv:1904.07248
Barrows, R. S., Comerford, J. M., Stern, D., & Assef, R. J. 2021, ApJ, 922, 179
Baum, W. A. 1957, AJ, 62, 6
Best, P. N., Kondapally, R., Williams, W. L., et al. 2023, MNRAS

[12] `https://lofar-surveys.org/`

[13] `http://www.star.bris.ac.uk/~mbt/topcat/`

[14] `https://jupyter.org`

[15] `https://pycaret.org`

[16] `https://pandas.pydata.org`

[17] `https://www.astropy.org`

[18] `https://betacal.github.io`

[19] `https://github.com/1313e/CMasher`

Bianchi, S., Chiaberge, M., Laor, A., et al. 2022, MNRAS, 516, 5775
Blecha, L., Snyder, G. F., Satyapal, S., & Ellison, S. L. 2018, MNRAS, 478, 3056
Bonaldi, A., An, T., Brüggen, M., et al. 2021, MNRAS, 500, 3821
Bonaldi, A., Bonato, M., Galluzzi, V., et al. 2019, MNRAS, 482, 2
Bonnarel, F., Fernique, P., Bienaymé, O., et al. 2000, A&AS, 143, 33
Bosman, S. E. I. 2022, The continuously updated webpage is hosted at: http://www.sarahbosman.co.uk/list_of_all_quasars
Bouwens, R., González-López, J., Aravena, M., et al. 2020, ApJ, 902, 112
Braun, R., Bonaldi, A., Bourke, T., Keane, E., & Wagg, J. 2019, arXiv e-prints, arXiv:1912.12699
Breiman, L. 2001, Machine Learning, 45, 5
Breiman, L. 2003, Statistics Department University of California Berkeley, CA, USA
Brier, G. W. 1950, Monthly Weather Review, 78, 1
Bröcker, J. & Smith, L. A. 2007, Weather and Forecasting, 22, 651
Brown, M. J. I., Duncan, K. J., Landt, H., et al. 2019, MNRAS, 489, 3351
Capetti, A., Brienza, M., Baldi, R. D., et al. 2020, A&A, 642, A107
Carilli, C. L., Furlanetto, S., Briggs, F., et al. 2004, New A Rev., 48, 1029
Carvajal, R., Bauer, F. E., Bouwens, R. J., et al. 2020, A&A, 633, A160
Carvajal, R., Matute, I., Afonso, J., et al. 2021, Galaxies, 9, 86
Casalicchio, G., Molnar, C., & Bischl, B. 2019, in Machine Learning and Knowledge Discovery in Databases, ed. M. Berlingerio, F. Bonchi, T. Gärtner, N. Hurley, & G. Ifrim (Cham: Springer International Publishing), 655–670
Chambers, K. C., Magnier, E. A., Metcalfe, N., et al. 2016, arXiv e-prints, arXiv:1612.05560
Chattopadhyay, A. K. 2017, Incomplete Data in Astrostatistics (American Cancer Society), 1–12
Chen, T. & Guestrin, C. 2016, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16 (New York, NY, USA: ACM), 785–794
Clarke, A. O., Scaife, A. M. M., Greenhalgh, R., & Griguta, V. 2020, A&A, 639, A84
Condon, J. J. 1992, ARA&A, 30, 575
Condon, J. J., Cotton, W. D., Greisen, E. W., et al. 1998, AJ, 115, 1693
Cover, T. & Hart, P. 1967, IEEE Transactions on Information Theory, 13, 21
Cramér, H. 1946, Mathematical methods of statistics (Princeton University Press Princeton), xvi, 575 p. ;
Cranmer, M. 2023, arXiv e-prints, arXiv:2305.01582
Cranmer, M., Sanchez-Gonzalez, A., Battaglia, P., et al. 2020, Advances in Neural Information Processing Systems, 33, 17429
Cunha, P. A. C. & Humphrey, A. 2022, A&A, 666, A87
Curran, S. J. 2022, MNRAS, 512, 2099
Curran, S. J., Moss, J. P., & Perrott, Y. C. 2022, MNRAS, 514, 1
Cutri, R. M., Skrutskie, M. F., van Dyk, S., et al. 2003a, 2MASS All Sky Catalog of point sources.
Cutri, R. M., Skrutskie, M. F., van Dyk, S., et al. 2003b, VizieR Online Data Catalog, II/246
Cutri, R. M., Wright, E. L., Conrow, T., et al. 2013, Explanatory Supplement to the AllWISE Data Release Products
Dahlen, T., Mobasher, B., Faber, S. M., et al. 2013, ApJ, 775, 93
Davies, L. J. M., Robotham, A. S. G., Driver, S. P., et al. 2018, MNRAS, 480, 768
Delhaize, J., Heywood, I., Prescott, M., et al. 2021, MNRAS, 501, 3833
della Ceca, R., Lamorani, G., Maccacaro, T., et al. 1994, ApJ, 430, 533
Desai, S. & Strachan, A. 2021, Scientific Reports, 11, 12761
Dey, B., Andrews, B. H., Newman, J. A., et al. 2022, MNRAS, 515, 5285
Dice, L. R. 1945, Ecology, 26, 297
Dorogush, A. V., Ershov, V., & Gulin, A. 2018, CoRR, abs/1810.11363 [arXiv:1810.11363]
Dorogush, A. V., Gulin, A., Gusev, G., et al. 2017, CoRR, abs/1706.09516 [arXiv:1706.09516]
Driver, S. P., Hill, D. T., Kelvin, L. S., et al. 2011, MNRAS, 413, 971
Duboue, P. 2020, The Art of Feature Engineering: Essentials for Machine Learning (Cambridge University Press)
Duncan, K. J., Sabater, J., Röttgering, H. J. A., et al. 2019, A&A, 622, A3
Euclid Collaboration, Bisigello, L., Conselice, C. J., et al. 2023a, MNRAS, 520, 3529
Euclid Collaboration, Humphrey, A., Bisigello, L., et al. 2023b, A&A, 671, A99
Fan, X., Banados, E., & Simcoe, R. A. 2022, arXiv e-prints, arXiv:2212.06907
Flesch, E. W. 2021, arXiv e-prints, arXiv:2105.12985
Flewelling, H. A., Magnier, E. A., Chambers, K. C., et al. 2020, ApJS, 251, 7
Friedman, J. H. 2001, The Annals of Statistics, 29, 1189
Friedman, J. H. 2002, Computational Statistics & Data Analysis, 38, 367, nonlinear Methods and Data Mining
Geurts, P., Ernst, D., & Wehenkel, L. 2006, Machine Learning, 63, 3
Gilda, S., Lower, S., & Narayanan, D. 2021, ApJ, 916, 43
Glahn, H. R. & Jorgensen, D. L. 1970, Monthly Weather Review, 98, 136
Gloudemans, A. J., Duncan, K. J., Röttgering, H. J. A., et al. 2021, A&A, 656, A137
Gloudemans, A. J., Duncan, K. J., Saxena, A., et al. 2022, A&A, 668, A27
Goebel, R., Chander, A., Holzinger, K., et al. 2018, in International cross-domain conference for machine learning and knowledge extraction, Springer (Springer International Publishing), 295–303
Gordon, Y. A., Boyce, M. M., O'Dea, C. P., et al. 2020, Research Notes of the American Astronomical Society, 4, 175
Gürkan, G., Hardcastle, M. J., Best, P. N., et al. 2019, A&A, 622, A11
Hardcastle, M. J. & Croston, J. H. 2020, New A Rev., 88, 101539
Head, T., Kumar, M., Nahrstaedt, H., Louppe, G., & Shcherbatyi, I. 2021, scikit-optimize/scikit-optimize
Heckman, T. M. & Best, P. N. 2014, ARA&A, 52, 589
Helfand, D. J., White, R. L., & Becker, R. H. 2015, ApJ, 801, 26
Helou, G., Soifer, B. T., & Rowan-Robinson, M. 1985, ApJ, 298, L7
Hernán-Caballero, A., Varela, J., López-Sanjuan, C., et al. 2021, A&A, 654, A101
Hickox, R. C. & Alexander, D. M. 2018, ARA&A, 56, 625
Hildebrandt, H., Arnouts, S., Capak, P., et al. 2010, A&A, 523, A31
Hill, G. J., Gebhardt, K., Komatsu, E., et al. 2008, in Astronomical Society of the Pacific Conference Series, Vol. 399, Panoramic Views of Galaxy Formation and Evolution, ed. T. Kodama, T. Yamada, & K. Aoki, 115
Hoaglin, D., Mosteller, F., Tukey, J., et al. 1983, Understanding Robust and Exploratory Data Analysis, Wiley Series in Probability and Statistics: Probability and Statistics Section Series (John Wiley & Sons)
Hodge, J. A., Becker, R. H., White, R. L., Richards, G. T., & Zeimann, G. R. 2011, AJ, 142, 3
Hopkins, A. M., Whiting, M. T., Seymour, N., et al. 2015, PASA, 32, e037
Hunter, J. D. 2007, Computing in Science & Engineering, 9, 90
Ilbert, O., Arnouts, S., McCracken, H. J., et al. 2006, A&A, 457, 841
Ilbert, O., Capak, P., Salvato, M., et al. 2009, ApJ, 690, 1236
Ilyas, I. F. & Rekatsinas, T. 2022, J. Data and Information Quality, 14
Inayoshi, K., Visbal, E., & Haiman, Z. 2020, ARA&A, 58, 27
Jarrett, T. H., Cluver, M. E., Magoulas, C., et al. 2017, ApJ, 836, 182
Jarvis, M., Taylor, R., Agudo, I., et al. 2016, in MeerKAT Science: On the Pathway to the SKA, 6
Jarvis, M. E., Harrison, C. M., Mainieri, V., et al. 2021, MNRAS, 503, 1780
Jensen, H., Datta, K. K., Mellema, G., et al. 2013, MNRAS, 435, 460
Jiang, L., Fan, X., Bian, F., et al. 2014, ApJS, 213, 12
Jiang, L., Fan, X., Ivezić, Ž., et al. 2007, ApJ, 656, 680
Johnson, N. & Leone, F. 1964, Statistics and Experimental Design in Engineering and the Physical Sciences, Vol. 2 (Wiley), 125
Jonas, J. & MeerKAT Team. 2016, in MeerKAT Science: On the Pathway to the SKA, 1
Ke, G., Meng, Q., Finley, T., et al. 2017, in Advances in Neural Information Processing Systems, ed. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett, Vol. 30 (Curran Associates, Inc.)
Kennicutt, Robert C., J., Hao, C.-N., Calzetti, D., et al. 2009, ApJ, 703, 1672
Kim, S. J., Lee, H. M., Matsuhara, H., et al. 2012, A&A, 548, A29
Kluyver, T., Ragan-Kelley, B., Pérez, F., et al. 2016, in Positioning and Power in Academic Publishing: Players, Agents and Agendas, ed. F. Loizides & B. Schmidt, IOS Press, 87 – 90
Kull, M., Filho, T. M. S., & Flach, P. 2017a, Electronic Journal of Statistics, 11, 5052
Kull, M., Filho, T. S., & Flach, P. 2017b, in Proceedings of Machine Learning Research, Vol. 54, Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, ed. A. Singh & J. Zhu (PMLR), 623–631
Lee, H. M., Kim, S. J., Im, M., et al. 2009, PASJ, 61, 375
Lehmer, B. D., Brandt, W. N., Alexander, D. M., et al. 2005, ApJS, 161, 21
Lichtenstein, S., Fischhoff, B., & Phillips, L. D. 1982, Calibration of probabilities: The state of the art to 1980, ed. D. Kahneman, P. Slovic, & A. Tversky (Cambridge University Press), 306334
Lima, E. V. R., Sodré, L., Bom, C. R., et al. 2022, Astronomy and Computing, 38, 100510
Liske, J., Baldry, I. K., Driver, S. P., et al. 2015, MNRAS, 452, 2087
Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. 2013, in Advances in Neural Information Processing Systems, ed. C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger, Vol. 26 (Curran Associates, Inc.)
Luken, K., Norris, R., Park, L., Wang, X., & Filipovi, M. 2022, Astronomy and Computing, 39, 100557
Luken, K. J., Norris, R. P., & Park, L. A. F. 2019, PASP, 131, 108003
Lundberg, S. M., Erion, G., Chen, H., et al. 2020, Nature Machine Intelligence, 2, 2522
Lundberg, S. M. & Lee, S.-I. 2017, in Advances in Neural Information Processing Systems 30, ed. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Curran Associates, Inc.), 4765–4774
Lyke, B. W., Higley, A. N., McLane, J. N., et al. 2020, ApJS, 250, 8
Ma, S. & Tourani, R. 2020, in Proceedings of Machine Learning Research, Vol. 127, Proceedings of the 2020 KDD Workshop on Causal Discovery (PMLR), 23–38
Macfarlane, C., Best, P. N., Sabater, J., et al. 2021, MNRAS, 506, 5888

Machado Poletti Valle, L. F., Avestruz, C., Barnes, D. J., et al. 2021, MNRAS, 507, 1468
Magliocchetti, M. 2022, A&A Rev., 30, 6
Mainzer, A., Bauer, J., Cutri, R. M., et al. 2014, ApJ, 792, 30
Mainzer, A., Bauer, J., Grav, T., et al. 2011, ApJ, 731, 53
Marocco, F., Eisenhardt, P. R. M., Fowler, J. W., et al. 2021, ApJS, 253, 8
Mateos, S., Alonso-Herrero, A., Carrera, F. J., et al. 2012, MNRAS, 426, 3271
Matthews, B. 1975, Biochimica et Biophysica Acta (BBA) - Protein Structure, 405, 442
McConnell, D., Hale, C. L., Lenc, E., et al. 2020, PASA, 37, e048
McGreer, I. D., Becker, R. H., Helfand, D. J., & White, R. L. 2006, ApJ, 652, 157
Miley, G. & De Breuck, C. 2008, A&A Rev., 15, 67
Mingo, B., Watson, M. G., Rosen, S. R., et al. 2016, MNRAS, 462, 2631
Norris, R. P., Hopkins, A. M., Afonso, J., et al. 2011, PASA, 28, 215
Norris, R. P., Salvato, M., Longo, G., et al. 2019, PASP, 131, 108004
Ochsenbein, F., Bauer, P., & Marcout, J. 2000, A&AS, 143, 23
Oliver, S., Rowan-Robinson, M., Alexander, D. M., et al. 2000, MNRAS, 316, 749
Pacifici, C., Iyer, K. G., Mobasher, B., et al. 2023, ApJ, 944, 141
Padovani, P. 1993, MNRAS, 263, 461
Padovani, P., Alexander, D. M., Assef, R. J., et al. 2017, A&A Rev., 25, 2
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825
Pensabene, A., Carniani, S., Perna, M., et al. 2020, A&A, 637, A84
Pierce, J. C. S., Tadhunter, C. N., Gordon, Y., et al. 2022, MNRAS, 510, 1163
Poisot, T. 2023, Methods in Ecology and Evolution, 14, 1333
Poliszczuk, A., Pollo, A., Małek, K., et al. 2021, A&A, 651, A108
Pouliasis, E. 2020, PhD thesis, IAASARS, National Observatory of Athens
Prandoni, I. & Seymour, N. 2015, in Advancing Astrophysics with the Square Kilometre Array (AASKA14), 67
Rasmussen, C. & Williams, C. 2006, Gaussian Processes for Machine Learning, Adaptative computation and machine learning series (University Press Group Limited)
Ratner, B. 2009, Journal of Targeting, Measurement and Analysis for Marketing, 17, 139
Reis, I., Baron, D., & Shahaf, S. 2019, AJ, 157, 16
Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. 2020, IEEE Access, 8, 42200
Ross, N. P. & Cross, N. J. G. 2020, MNRAS, 494, 789
Saarela, M. & Jauhiainen, S. 2021, SN Applied Sciences, 3, 272
Salvato, M., Ilbert, O., & Hoyle, B. 2019, Nature Astronomy, 3, 212
Samuel, A. L. 1959, IBM Journal of Research and Development, 3, 210
Sánchez-Sáez, P., Reyes, I., Valenzuela, C., et al. 2021, AJ, 161, 141
Scoville, N., Aussel, H., Brusa, M., et al. 2007, ApJS, 172, 1
Shapley, L. S. 1953, A Value for n-Person Games, Vol. 1 (Princeton University Press), 307–318
Shimwell, T. W., Tasse, C., Hardcastle, M. J., et al. 2019, A&A, 622, A1
Shobhana, D., Norris, R. P., Filipović, M. D., et al. 2023, MNRAS, 519, 4902
Shy, S., Tak, H., Feigelson, E. D., Timlin, J. D., & Babu, G. J. 2022, AJ, 164, 6
Silva Filho, T., Song, H., Perello-Nieto, M., et al. 2021, arXiv e-prints, arXiv:2112.10327
Singh, D. & Singh, B. 2020, Applied Soft Computing, 97, 105524
Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, AJ, 131, 1163
Sørenson, T. 1948, A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content, Biologiske skrifter (I kommission hos E. Munksgaard)
Stern, D., Assef, R. J., Benford, D. J., et al. 2012, ApJ, 753, 30
Stone, M. 1974, Journal of the Royal Statistical Society: Series B (Methodological), 36, 111
Storchi-Bergmann, T. & Schnorr-Müller, A. 2019, Nature Astronomy, 3, 48
Taylor, M. B. 2005, in Astronomical Society of the Pacific Conference Series, Vol. 347, Astronomical Data Analysis Software and Systems XIV, ed. P. Shopbell, M. Britton, & R. Ebert, 29
Thomas, N., Davé, R., Jarvis, M. J., & Anglés-Alcázar, D. 2021, MNRAS, 503, 3492
Thorne, J. E., Robotham, A. S. G., Davies, L. J. M., et al. 2022, MNRAS, 509, 4940
Van Calster, B., McLernon, D. J., van Smeden, M., et al. 2019, BMC Medicine, 17, 230
van der Velden, E. 2020, The Journal of Open Source Software, 5, 2004
van Haarlem, M. P., Wise, M. W., Gunst, A. W., et al. 2013, A&A, 556, A2
van Rijsbergen, C. J. 1979, Information Retrieval, 2nd edn. (USA: Butterworth-Heinemann)
Vanschoren, J. 2019, Meta-Learning, ed. F. Hutter, L. Kotthoff, & J. Vanschoren (Cham: Springer International Publishing), 35–61
Villaescusa-Navarro, F., Anglés-Alcázar, D., Genel, S., et al. 2021, ApJ, 915, 71
Villar-Martín, M., Emonts, B., Cabrera Lavers, A., et al. 2017, MNRAS, 472, 4659
Walcher, J., Groves, B., Budavári, T., & Dale, D. 2011, Ap&SS, 331, 1
Werner, M. W., Roellig, T. L., Low, F. J., et al. 2004, ApJS, 154, 1
Wes McKinney. 2010, in Proceedings of the 9th Python in Science Conference, ed. Stéfan van der Walt & Jarrod Millman, 56 – 61
Williams, W. L., Calistro Rivera, G., Best, P. N., et al. 2018, MNRAS, 475, 3429
Wolpert, D. H. 1992, Neural Networks, 5, 241
Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868
Wu, C., Wong, O. I., Rudnick, L., et al. 2019, MNRAS, 482, 1211
Yang, J. 2021, arXiv e-prints, arXiv:2109.09847
Yeo, I. & Johnson, R. A. 2000, Biometrika, 87, 954
Yerushalmy, J. 1947, Public Health Reports (1896-1970), 62, 1432
York, D. G., Adelman, J., Anderson, John E., J., et al. 2000, AJ, 120, 1579
Yule, G. U. 1912, Journal of the Royal Statistical Society, 75, 579
Zheng, A. & Casari, A. 2018, Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists (O'Reilly)
Zitlau, R., Hoyle, B., Paech, K., et al. 2016, MNRAS, 460, 3152

Table A.1. Names of columns or features used in the code and what they represent.

Photometry measurements (magnitudes and fluxes)					
Code name	Feature	Code name	Feature	Code name	Feature
`gmag`	g (PS1)	`ymag`	y (PS1)	`W1mproPM`	W1 (CW)
`rmag`	r (PS1)	`Jmag`	J (2M)	`W1mproPM`	W2 (CW)
`imag`	i (PS1)	`Hmag`	H (2M)	`W3mag`	W3 (AW)
`zmag`	z (PS1)	`Kmag`	Ks (2M)	`W4mag`	W4 (AW)
Colours					
66 colours from all combinations of non-radio magnitudes. A sub-sample of them is shown.					
`g_r`	g - r (PS1)	...	...	`W2_W3`	W2 (CW) - W3 (AW)
`g_i`	g - i (PS1)	...	...	`W2_W4`	W2 (CW) - W4 (AW)
`g_z`	g - z (PS1)	...	...	`W3_W4`	W3 - W4 (AW)
Categorical flags					
Code name	Feature				
`band_num`	Number of bands with measurements				
Boolean flags					
Code name	Feature	Code name	Feature		
`class`	AGN or galaxy	`radio_detect`	Detection in, at least, one radio band.		
Redshift					
Code name	Feature				
`z`	Spectroscopic redshift				
Outputs of base models					
Code name	Feature	Code name	Feature	Code name	Feature
`XGBoost`	XGBoost	`ET`	Extra Trees	`GBR`	Gradient Boosting Regressor
`CatBoost`	CatBoost	`GBC`	Gradient Boosting Classifier		
`RF`	Random Forest				

Table B.1. Density of detected sources (in units of sources per square degree) per band and field.

HETDEX Field							
Band	Density (deg^{-2})	Band	Density (deg^{-2})	Band	Density (deg^{-2})	Band	Density (deg^{-2})
g	6 380.66	z	10 331.93	H	1 335.55	W2	35 700.18
r	9 304.58	y	6 735.97	Ks	1 335.55	W3	14 045.08
i	11 242.35	J	1 335.55	W1	35 700.18	W4	14 044.78
S82 Field							
Band	Density (deg^{-2})	Band	Density (deg^{-2})	Band	Density (deg^{-2})	Band	Density (deg^{-2})
g	8 249.04	z	13 214.70	H	2 330.92	W2	39 025.05
r	12 962.35	y	9 226.45	Ks	2 330.92	W3	15 393.12
i	14 507.01	J	2 330.92	W1	39 025.01	W4	15 472.75

Appendix A: Column names in this study

Table A.1 presents the names (and what they represent) of the features, used in throughout this work. This information can be read in combination with the columns presented in Appendix G.

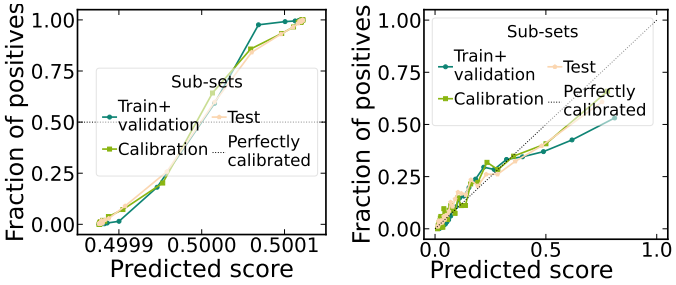
Appendix B: Non-imputed magnitudes

The number of valid measurements in Fig. 1 for each field and band can be used to determine the relative difference of density of sources between both fields. This density can be obtained by dividing the number of valid measurements over the effective area of each field (Sect. 2). Table B.1 shows these densities.

Appendix C: From Scores to Probabilities

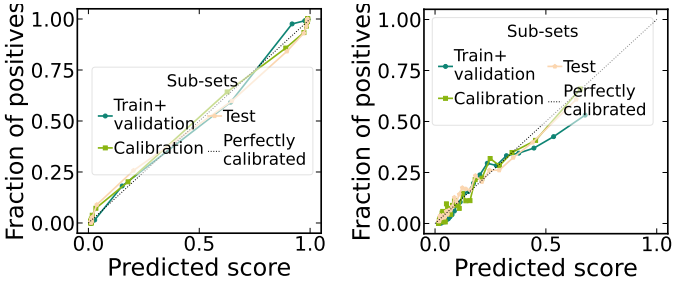
In general, classifiers deliver scores in the range [0, 1], which could be associated to the probability of a studied source being part of the relevant class (in our work, AGN or radio detectable). The classifier uses a threshold above which, any predicted element would be considered a positive instance.

With the exception of few algorithms (including the family of logistic regressions), scores from classifiers cannot be directly used as probabilities. As a consequence of this inability, such values cannot be compared from one type of model to some other and can not be combined to obtain a joint score. Therefore, in order to retrieve joint scores and treat them as probabilities, scores



(a) AGN-galaxy (b) Radio detection

Fig. C.1. Reliability curves for uncalibrated classifiers. Each line represents the calibration curve for each subset in HETDEX field. Data has been binned and each bin (represented by the points) has the same number of elements per curve. Dashed line represents a perfectly calibrated model. (a) AGN-galaxy classification model. (b) Radio detection model.



(a) AGN-galaxy (b) Radio detection

Fig. C.2. Reliability curves for calibrated classifiers. (a) AGN-galaxy classification model. (b) Radio detection model. Details as in Fig. C.1.

(and, by extension, the classifiers) need to be calibrated. This calibration means that, when taking all predictions with a probability P of being of a class, a fraction P of them really belong to that class (e.g. Lichtenstein et al. 1982; Silva Filho et al. 2021).

Calibration of these scores can be done by applying a transformation to their values. For our work, we will apply a Beta transformation. It allows to re-distribute the scores of a classifier allowing them to get closer to the definition of probability (Kull et al. 2017a,b). Calibration steps in our workflow have been applied using the Python package `betacal`. In the case of the radio detection model, the new scores have a wider range than the original, uncalibrated scores.

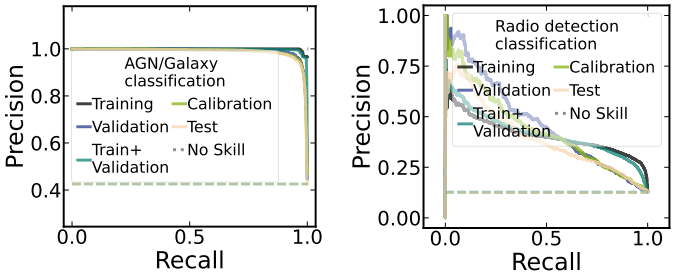
When obtaining the BSS values for both classification, the AGN-Galaxy classifier has a score of BSS = −0.002, demonstrating that no major changes were applied to the distribution of scores. For the radio detection classifier, the score is BSS = −0.434. Even though the BSS value is slightly negative for the AGN-Galaxy classifier, we keep it since its range of values now can be compared and combined with additional probabilities. In the case of the radio detection classifier, the BSS shows a degradation of the calibration, but we will keep the calibrated model given that it provides, overall, better values for the remaining metrics. This effect can be seen, more strongly, with recall.

Calibration (or reliability) plots show how well calibrated the predicted scores of a classifier are by displaying the fraction of sources that are part of a given class as a function of the predicted probability. A perfectly calibrated classifier would have all its prediction lying in the x=y line. The magnitude of the deviations from that line give information of the miscalibration a model has (see, for instance, Bröcker & Smith 2007; Van Cal-

Table D.1. Hyper-parameters values for meta-learners after tuning.

AGN-Galaxy model (CatBoost)			
Parameter	Value	Parameter	Value
learning_rate	0.0075	random_strength	0.1
depth	6	l2_leaf_reg	10
Radio detection model (GradientBoosting)			
Parameter	Value	Parameter	Value
n_estimators	187	min_samples_leaf	2
learning_rate	0.0560	max_depth	9
subsample	0.3387	max_features	0.5248
min_samples_split	5		
Redshift prediction model (ET)			
Parameter	Value	Parameter	Value
n_estimators	100	criterion	mae
max_depth	None	min_samples_split	2
max_features	auto	min_samples_leaf	1
bootstrap	False		

Notes. This table shows the parameters which were subject to tuning. Remaining hyper-parameters used their default values as defined by their developers.



(a) AGN-galaxy (b) Radio detection

Fig. E.1. Precision-Recall curves for the (a) AGN-Galaxy and (b) radio detection classification models.

ster et al. 2019). In Fig. C.1, we present the reliability curves for the uncalibrated classifiers and, in Fig. C.2, for their calibrated versions.

Appendix D: Meta-learners hyper-parameters

In Table D.1, we present the optimised hyper-parameters from our meta-learners. For all three instances of modelling (AGN-Galaxy, radio detection, and redshift), hyper-parameters were optimized using the `SkoptSearch` algorithm embedded in the package `tune-sklearn`[20] (`v0.4.1`; Head et al. 2021), which implements a bayesian search in the hyper-parameter space.

Appendix E: PR-curve threshold optimisation

By maximising the recall (Eq. 4), we improve the number of recovered elements in each classifier**. This can be** done by decreasing the threshold by which a source is classified as a positive instances. Setting this threshold to its minimum, 0.0, would increase the recall. But every source would be predicted to be an AGN or detected on the radio regardless of their properties.

One statistical tool designed to optimise the classification threshold taking into account the **overall model performance** is the Precision-Recall (PR) curve. It can help to understand the behaviour of a classifier as a function of its threshold. Both quantities, precision (Eq. 5) and recall, show an inverse correlation, and both depend on the selected threshold. Thus, they can be used to retrieve the score value for which both quantities are balanced. This optimisation is done by finding the threshold that maximises the F_β score (Eq. 2). This operation will be performed over the union of training and validation sets, which have been used to create and train each model. PR curves for all subsets used in our classification models are shown in Fig. E.1.

[20] `https://github.com/ray-project/tune-sklearn`

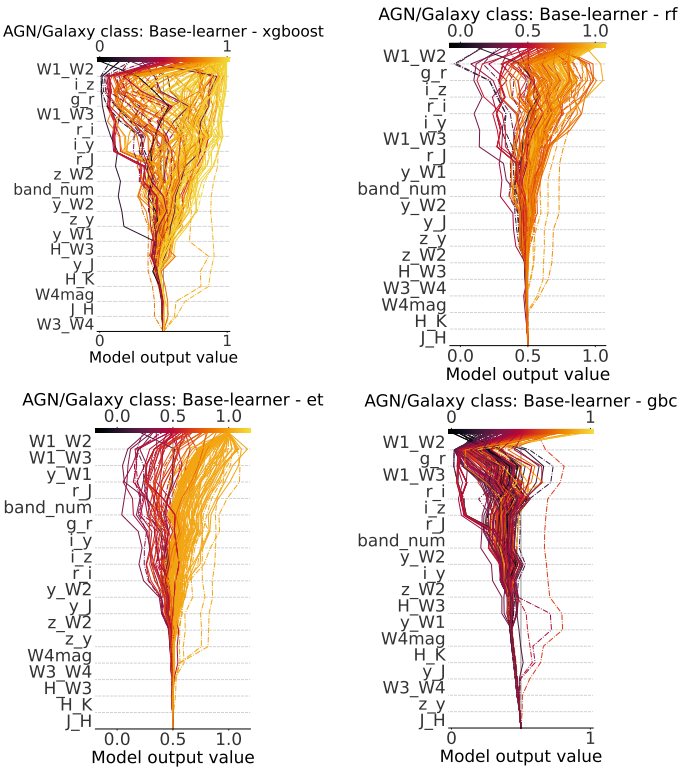


Fig. F.1. SHAP decision plots for base AGN-Galaxy algorithms. Details as described in Figs. 13. Starting point of predictions is the naive classification threshold. From left to right and from top to bottom, each panel shows the results from `XGBoost`, `RF`, `ET`, and `GBC`.

Appendix F: SHAP values for base models

Figures F.1, F.3, and F.2 show the decision plots for each base learners used in the prediction models of our pipeline (Sect. 5.3.2). For the classification algorithms, the starting point of their predictions corresponds to the naive threshold (0.5) since no threshold optimisation was applied to them and only the scores are included in the stacking step, not the final **probabilities**. In the case of the redshift predictors, decision plots start at the value z=0, as presented for the meta-learner.

Appendix G: Prediction results for radio AGN

The columns shown in the prediction results **for sources in both HETDEX and S82 fields** are described **as follows**.

- `ID`: Internal identification number.
- `RA_ICRS`: Right Ascension (in degrees) of source in CW.
- `DE_ICRS`: Declination (in degrees) of source in CW.
- **`Name`: Name of the source as it appears in the CW catalogue.**
- `band_num`: Number of non-radio bands with a valid measurement per source (cf. Sect. 2.2).

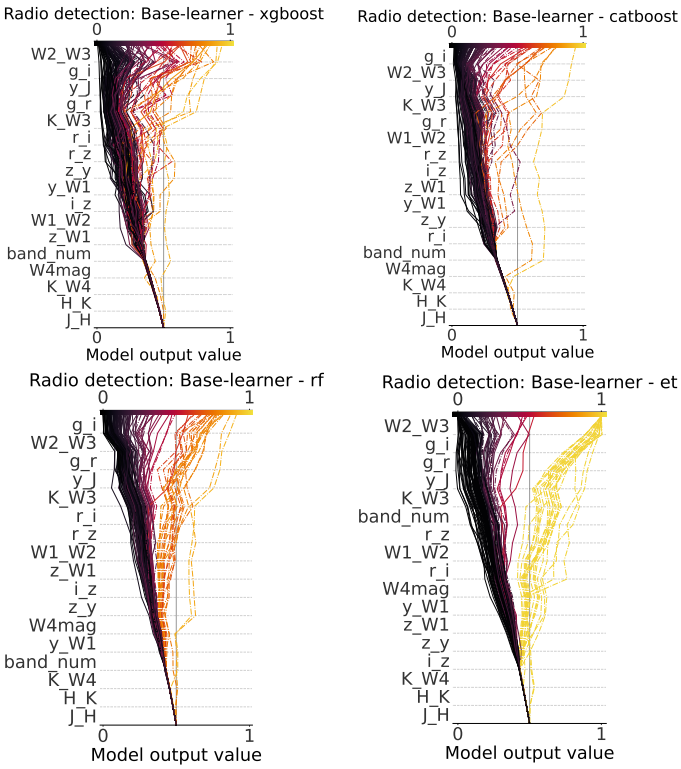


Fig. F.2. SHAP decision plots from base radio algorithms. Details as Figs. 13 and F.1. Each panel with results for `XGBoost`, `CatBoost`, `RF`, and `ET`.

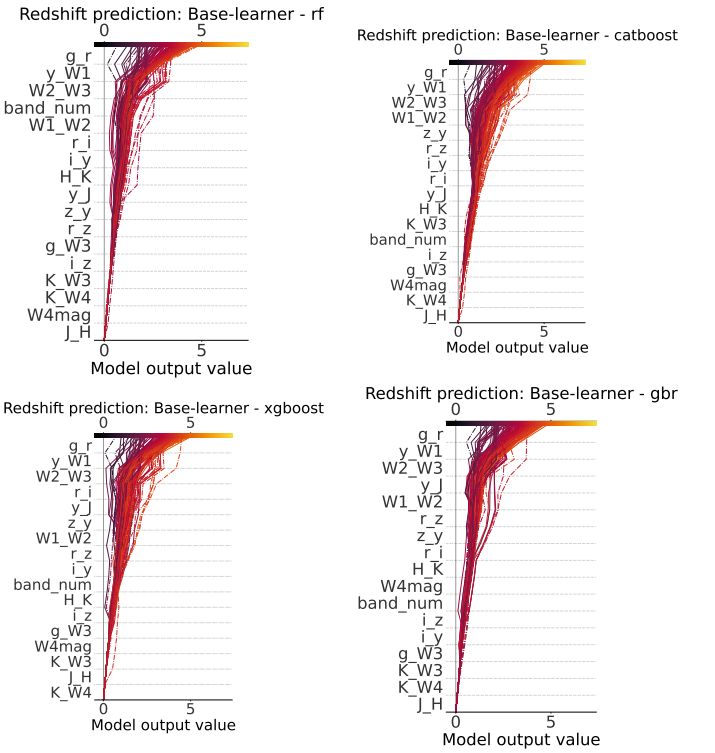


Fig. F.3. SHAP decision plots from base z algorithms. Details as in Fig 13. Each panel shows results for `ET`, `CatBoost`, `XGBoost`, and `GBR`.

- `class`: 1 if a source is a confirmed AGN by the MQC. `0` if it has been spectroscopically confirmed as a galaxy in SDSS DR16. Sources with no value do not have a spectroscopic classification in this catalogue.
- **`Sint_LOFAR` (or `Fint_VLAS82`): Imputed integrated flux (in mJy) of source from LOFAR or VLAS82.**
- **`Sint_LOFAR_non_imp` (or `Fint_VLAS82_non_imp`): Non imputed integrated flux (in mJy) of source from LOFAR or VLAS82.**
- **`W1mproPM`: Imputed W1 magnitude of source.**
- **`W2mproPM`: Imputed W2 magnitude of source.**
- **`gmag`: Imputed g magnitude of source.**
- **`rmag`: Imputed r magnitude of source.**
- **`imag`: Imputed i magnitude of source.**
- **`zmag`: Imputed z magnitude of source.**
- **`ymag`: Imputed y magnitude of source.**
- **`W3mag`: Imputed W3 magnitude of source.**
- **`W4mag`: Imputed W4 magnitude of source.**
- **`Jmag`: Imputed J magnitude of source.**
- **`Hmag`: Imputed H magnitude of source.**
- **`Kmag`: Imputed Ks magnitude of source.**
- `Score_AGN`: Score from the meta AGN-Galaxy classifier for a prediction to be an AGN.
- `Prob_AGN`: Probability from the calibrated meta AGN-Galaxy classifier for a prediction to be an AGN.
- `LOFAR_detect`: `1` if a source has been detected on the LoTSS survey or in their analogue surveys for fields different to HETDEX (see Sect. 2.1). `0` otherwise.
- `Score_radio_AGN`: Score from the meta radio detection model for a prediction to be detected in the radio.
- `Prob_radio_AGN`: Probability, from the calibrated radio detection model for a prediction to be detected in the radio.
- **`radio_AGN`: `class` × `LOFAR_detect`. `1` if a source is an AGN and has been detected in the radio. `0` otherwise.**
- `Score_rAGN`: `Score_AGN` × `Score_radio`. Score of a source for it to be an AGN detected in the radio.
- `Prob_rAGN`: `Prob_AGN` × `Prob_radio`. Probability of a source for it to be an AGN detected in the radio.
- `Z`: Spectroscopic redshift as listed by the MQC (if available).
- `pred_Z`: Redshift value predicted by our model.