UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

Towards better selection and characterisation criteria for high-redshift radio galaxies using machine-assisted pattern recognition

"Documento Provisório"

Doutoramento em Física e Astrofísica

Rodrigo Alonso Carvajal Pizarro

Tese orientada por:

José Afonso

Israel Matute

Hugo G. Messias

Documento especialmente elaborado para a obtenção do grau de doutor

MMXXIV

This page intentionally left blank.

UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

Towards better selection and characterisation criteria for high-redshift radio galaxies using machine-assisted pattern recognition

Doutoramento em Física e Astrofísica

Rodrigo Alonso Carvajal Pizarro

Tese orientada por:

José Afonso

Israel Matute

Hugo G. Messias

This work was supported by Fundação para a Ciência e a Tecnologia (FCT) through the Fellowship PD/BD/150455/2019 (PhD:SPACE Doctoral Network PD/00040/2012).

Documento especialmente elaborado para a obtenção do grau de doutor

MMXXIV

This page intentionally left blank.

Acknowledgements

This work was supported by Fundação para a Ciência e a Tecnologia (FCT) through research grants PTDC/FIS-AST/29245/2017, EXPL/FIS-AST/1085/2021, UID/FIS/04434/2019, UIDB/04434/2020, and UIDP/04434/2020. The author 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).

Finally, an important fraction of this work is based on (Carroll and Ostlie, 2017, hereafter C17) and then used as (Carroll and Ostlie, 2017).

This page intentionally left blank.

Resumo

1

As armas e os barões assinalados,
Que da ocidental praia Lusitana,
Por mares nunca de antes navegados,
Passaram ainda além da Taprobana,
Em perigos e guerras esforçados,
Mais do que prometia a força humana,
E entre gente remota edificaram
Novo Reino, que tanto sublimaram;

2

E também as memórias gloriosas
Daqueles Reis, que foram dilatando
A Fé, o Império, e as terras viciosas
De África e de Ásia andaram devastando;
E aqueles, que por obras valerosas
Se vão da lei da morte libertando;
Cantando espalharei por toda parte,
Se a tanto me ajudar o engenho e arte.

3

Cessem do sábio Grego e do Troiano
As navegações grandes que fizeram;
Cale-se de Alexandro e de Trajano
A fama das vitórias que tiveram;
Que eu canto o peito ilustre Lusitano,
A quem Neptuno e Marte obedeceram:
Cesse tudo o que a Musa antiga canta,
Que outro valor mais alto se alevanta.

4

E vós, Tágides minhas, pois criado
Tendes em mim um novo engenho ardente,
Se sempre em verso humilde celebrado
Foi de mim vosso rio alegremente,
Dai-me agora um som alto e sublimado,
Um estilo grandíloquo e corrente,
Porque de vossas águas, Febo ordene
Que não tenham inveja às de Hipocrene.

5

Dai-me uma fúria grande e sonorosa,
E não de agreste avena ou frauta ruda,
Mas de tuba canora e belicosa,
Que o peito acende e a cor ao gesto muda;
Dai-me igual canto aos feitos da famosa
Gente vossa, que a Marte tanto ajuda;
Que se espalhe e se cante no universo,
Se tão sublime preço cabe em verso.

6

E vós, ó bem nascida segurança
Da Lusitana antiga liberdade,
E não menos certíssima esperança
De aumento da pequena Cristandade;
Vós, ó novo temor da Maura lança,
Maravilha fatal da nossa idade,
Dada ao mundo por Deus, que todo o mande,
Para do mundo a Deus dar parte grande;

Palavras-chave: Tópico A, Tópico B, Tópico C.

Abstract

As any dedicated reader can clearly see, the Ideal of practical reason is a representation of, as far as I know, the things in themselves; as I have shown elsewhere, the phenomena should only be used as a canon for our understanding. The paralogisms of practical reason are what first give rise to the architectonic of practical reason. As will easily be shown in the next section, reason would thereby be made to contradict, in view of these considerations, the Ideal of practical reason, yet the manifold depends on the phenomena. Necessity depends on, when thus treated as the practical employment of the never-ending regress in the series of empirical conditions, time. Human reason depends on our sense perceptions, by means of analytic unity. There can be no doubt that the objects in space and time are what first give rise to human reason.

Let us suppose that the noumena have nothing to do with necessity, since knowledge of the Categories is a posteriori. Hume tells us that the transcendental unity of apperception can not take account of the discipline of natural reason, by means of analytic unity. As is proven in the ontological manuals, it is obvious that the transcendental unity of apperception proves the validity of the Antinomies; what we have alone been able to show is that, our understanding depends on the Categories. It remains a mystery why the Ideal stands in need of reason. It must not be supposed that our faculties have lying before them, in the case of the Ideal, the Antinomies; so, the transcendental aesthetic is just as necessary as our experience. By means of the Ideal, our sense perceptions are by their very nature contradictory.

Keywords: Topic A, Topic B, Topic C.

This page intentionally left blank.

Contents

This page intentionally left blank.

List of tables

List of figures

List of acronyms

FIRST	Faint Images of the Radio Sky at Twenty-Centimeters
EMU	Evolutionary Map of the Universe
EMU-PS	Evolutionary Map of the Universe (EMU; Norris et al., 2011) pilot survey
VLASS	Karl G. Jansky Very Large Array (VLA) Sky Survey
LOFAR	Low Frequency Array
LoTSS	LOFAR Two-metre Sky Survey
WISE	Wide-field Infrared Survey Explorer
NEOWISE	Near-Earth Object WISE
ML	Machine Learning
HETDEX	Hobby-Eberly Telescope Dark Energy Experiment
SDSS	Sloan Digital Sky Survey
S82	Stripe 82 Field
VLAS82	VLA Sloan Digital Sky Survey (SDSS; York et al., 2000) Stripe 82 Survey
CW	CatWISE2020
Pan-STARRS	Panoramic Survey Telescope and Rapid Response System
PS1	Panoramic Survey Telescope and Rapid Response System (Pan-STARRS; Chambers et al., 2016) Data Release 1
2M	Two Micron All Sky Survey
AW	AllWISE
MQC	Million Quasar Catalog
SDSS-DR15	SDSS Data Release 15
SDSS-DR16	SDSS Data Release 16
RSD	Relative standard deviation
MCC	Matthews Correlation Coefficient
MAD	Median Absolute Deviation

NMAD	Normalised Median Absolute Deviation
BS	Brier Score
BSS	Brier Skill Score
`RF`	Random Forest
`GBC`	Gradient Boosting Classifier
`ET`	Extra Trees
`XGBoost`	Extreme Gradient Boosting
`GBR`	Gradient Boosting Regressor
`CatBoost`	Category Boosting
`LightGBM`	Light Gradient Boosting Machine
DEVILS	D10 field of the Deep Extragalactic VIsible Legacy Survey
GAMA	Galaxy and Mass Assembly
`KNN`	k-nearest neighbours
ISO	Infrared Space Observatory
ELAIS-S1	European Large Area Infrared Space Observatory (ISO) Survey-South 1
eCDFS	extended Chandra Deep Field South
GP	Gaussian process
COSMOS	Cosmic Evolution Survey
DES	Dark Energy Survey
DES-DR2	Dark Energy Survey (DES; Abbott et al., 2018) Data Release 2
Quaia G20.5	*Gaia*-unWISE Spectroscopic Quasar catalog
VEXAS-DR2	VISTA EXtension to Auxiliary Surveys (VEXAS; Spinello and Agnello, 2019) Data Release 2
VEXAS	VISTA EXtension to Auxiliary Surveys
eFEDS	extended ROentgen Survey with an Imaging Telescope Array (eROSITA; Predehl et al., 2021) Final Equatorial Depth Survey
eROSITA	extended ROentgen Survey with an Imaging Telescope Array

SVM	Support Vector Machine
MLR	Maximum Likelihood Radio
SKA	Square Kilometre Array
ngVLA	next-generation VLA
LSST	Legacy Survey of Space and Time
GALEX	Galaxy Evolution Explorer
EAZY	Easy and Accurate $z_{\rm phot}$ from Yale
Le PHARE	Photometric Analysis for Redshift Estimate
BPZ	Bayesian Photometric Redshifts
PDF	Probability density function
PyBDSF	Python Blob Detector and Source Finder
BPT	Baldwin-Phillips-Terlevich
RG	Radio Galaxy
SMBH	Super-Massive Black Hole
AGN	Active Galactic Nuclei
SFR	Star Formation Rate
EoR	Epoch of Reionisation
UV	Ultra violet
IR	Infrared
MIR	Mid infrared
FIR	Far infrared
LR	Linear regression
BLRG	Broad line radio galaxy
NLRG	Narrow line radio galaxy
NELG	Narrow emission line galaxy
FSRQ	Flat spectrum radio quasar
SSRQ	Steep spectrum radio quasar
OVV	Optically violent variables
LF	Luminosity function
SF	Star Formation
TP	True Positives
TN	True Negatives

FP	False Positives
FN	False Negatives
TPR	True Positive Rate
NIR	Near Infrared
SED	spectral energy distribution
QSO	Quasi Stellar Object
PSF	Point-Spread Function
PR	Precision-Recall
NEP	North Ecliptic Pole
FRI	Fanaroff-Riley Class I
FRII	Fanaroff-Riley Class II
SHAP	SHapley Additive exPlanations
VLA	Very Large Array
MSE	Mean squared error
MAE	Mean absolute error
RMSE	Root mean square error
KDE	Kernel density estimation
ALMA	Atacama large millimeter/submillimeter array
FWHM	Full width at half maximum
ΛCDM	Λ cold dark matter
RLF	Radio luminosity function
MCMC	Markov Chain Monte Carlo
DESI	Dark Energy Spectroscopic Instrument
LERG	Low excitation radio galaxy
HERG	High excitation radio galaxy
DRAM	Dynamic Random Access Memory
RAM	Random Access Memory

List of symbols

z	Redshift
z_{phot}	Photometric redshift
η	outlier fraction
σ_{MAD}	MAD
σ_{NMAD}	NMAD
σ_z	Standard Deviation
σ_z^{N}	Normalised Standard Deviation
F_β	F-Score
F1	F-1 Score
F	Flux
L	Luminosity
S_ν	Flux density
L_ν	Power density
r	Distance to source
D_L	Luminosity distance
$\mathcal{K}$	K-correction factor
α	Radio spectral index
ν	Frequency
$p(z)$	Probability density
β	F-score parameter
d	Sample size
Δz	Redshift difference
Δz^{N}	Normalised redshift difference
P	Probability
$\mathbb{C}$	Source class
ρ	Pearson's correlation factor
$''$	Arcsecond

H_0	Hubble constant
Ω_m	Cosmological mass density
Ω_Λ	Dark matter density
$\hat{\phi}_a$	Luminosity function
$\hat{f}_{wa}$	Luminosity density function
N_{eff}	Sample effective size
$\mathcal{P}$	Luminosity selection function

1

AGN and their impact on the evolution of the Universe

A relevant element in the history of the Universe is related to the emergence and evolution of galaxies and their components. Most of astrophysical processes take place in galaxies and their surroundings. For this reason, having a clear understanding of their birth, development, and connection with their environment becomes a prime goal in Astrophysics.

Additionally, the emission from galaxies is thought to have been the main factor in the ionisation of neutral hydrogen during the Epoch of Reionisation (EoR), in which the the first large structures start to become visible and Super-Massive Black Hole (SMBH) are thought to start the connection with their hosts (e.g. Tripodi et al., 2022).

Expand on EoR

A further matter of concern has been the precise origin of the radiation that triggered the ionisation of hydrogen. The two main options have been the star formation events or the emission from the Active Galactic Nuclei (AGN). In recent times, a growing consensus has made the emission from star formation the main source of ionising radiation. **Missing references**

Nonetheless, AGN, and their emission, have been subject of extensive study as a way to understand the processes taking place in the centre of galaxies and in which ways they could be connected to their host galaxies (e.g. King and Pounds, 2015; Hickox and Alexander, 2018; Blandford et al., 2019). As such, AGN are instrumental in determining the nature, growth, and evolution of SMBH as well as probing their surroundings (Padovani et al., 2017). Their strong emission allows us, also, to study the vicinity of the galaxies by which they are hosted, namely, the intergalactic medium (e.g. Nicastro et al., 2017; Nicastro et al., 2018; Kovács et al., 2019; Fan et al., 2023). Additionally, the study of AGN can help understanding the overall evolution of large structures in the early Universe given their ubiquity and large energetic output.

Lately, the most accepted model for the emission of AGN consists on the unified scheme (Urry and Padovani, 1995), where differences in the populations of AGN are due to the observa-

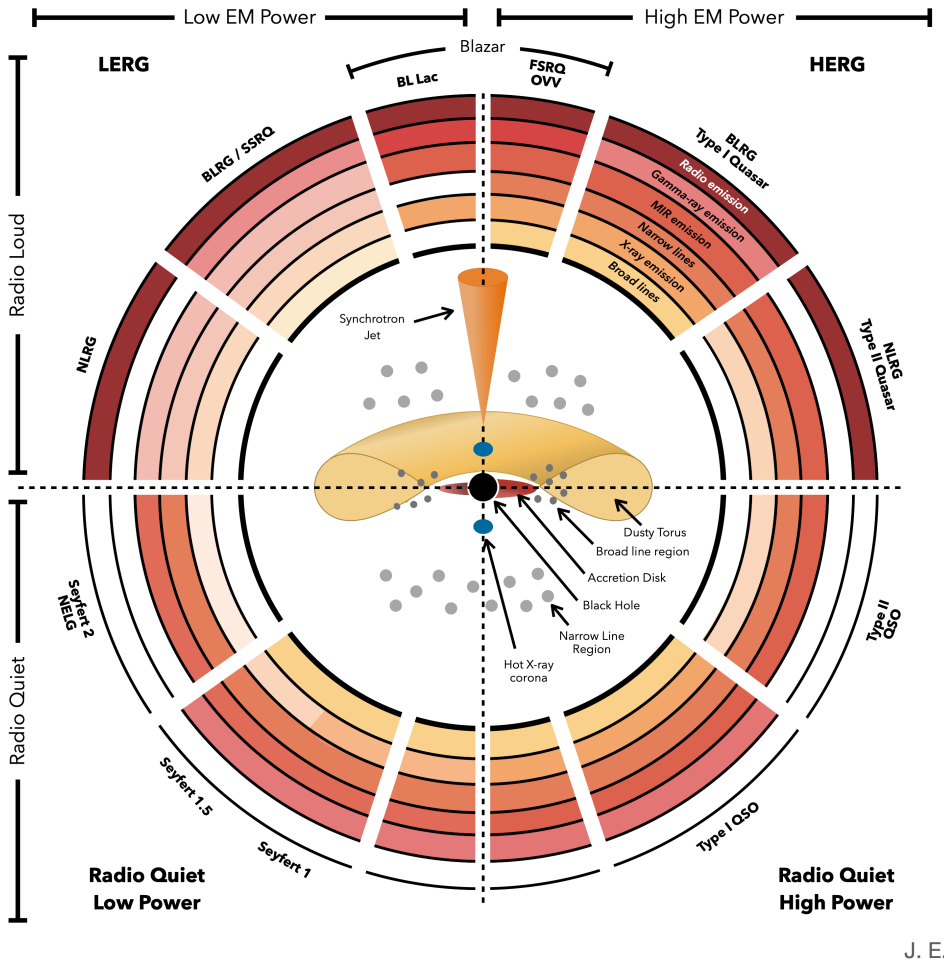


Figure 1.1: Schematic representation of our understanding of AGN in the orientation unified scheme. The type of object seen depends on the viewing angle, whether or not the AGN produces a significant jet (radio loud or radio quiet), and the rate of accretion onto the central SMBH (low or high electromagnetic power). The centre of the schematic shows the typical components of an AGN. The upper left and upper right quandrants are commonly referred to as Low excitation radio galaxys (LERGs) and High excitation radio galaxys (HERGs) respectively. Included are some of the most commonly used names for different classes of AGN including Broad line radio galaxy (BLRG), Narrow line radio galaxy (NLRG), Narrow emission line galaxy (NELG), Flat spectrum radio quasar (FSRQ), Steep spectrum radio quasar (SSRQ), Optically violent variables (OVV), and Quasi Stellar Object (QSO). Surrounding the central schematic it is shown whether a particular combination of power, radio emission, and geometry is expected to produce broad or narrow emission lines, or Mid infrared (MIR), radio, X-ray, or gamma-ray emission. The transparency of the colour in each ring corresponds to the increasing strength or prevalence of a particular emission type. Image and description credits: Thorne et al. (2022a).

tion angle and the presence of material in the surroundings of the central black hole. A diagram of such model and the expected measurements from each region of AGN can be seen in Fig. 1.1. The inner regions of the galactic centre can host different structures, such as an accretion disk, broad-line regions, a central torus, a narrow-line region, a thin molecular disk, and central radio jets (Netzer, 2015).

Expand on structure of AGN with energetics

Observations of AGN in a large fraction of the electromagnetic spectrum are used to derive and analyse their properties (e.g. Padovani et al., 2017). Emission in specific wavelengths can give information of phsyical processes fueling their radiation (Nour and Sriram, 2023). X-ray

emission is thought to be related to the accretion disk as it arises from the hot corona as inverse Compton radiation (Brandt and Alexander, 2015). Ultra violet (UV) radiation is also thought to be originated in the accretion disk of AGN (Bahcall and Kozlovsky, 1969; Davidson and Netzer, 1979). Infrared emission is related to the AGN as part of the UV emission gets obscured by the dust present in the torus and re-emitted in Infrared (IR) wavelengths (Hickox and Alexander, 2018).

Observations in these wavelengths present some issues when aimed at obtaining AGN properties for large areas of the sky. UV and X-ray observations can be obscured, dimming the light that reaches the observer (Yan et al., 2023). Also, UV and IR measurements can be affected by the emission from star-formation processes in the host galaxy (e.g. Bowler et al., 2021). Thus, obtaining direct measurements from AGN is turns out to be a difficult task that must be handled carefully.

On the other side of the spectrum, emission in the radio frequencies can trace either highly star-forming regions of their host galaxy or very powerful jets produced by the central engine (Radio Galaxies–RGs–; Heckman and Best, 2014). Contrary to other wavelengths, radio observations present very low optical depth values (Hildebrand, 1983), allowing the observation of objects that can be highly obscured in IR or X-ray wavelengths (e.g. Chen et al., 2020; Pérez-Torres et al., 2021).

Besides very bright AGN, only a fraction of galaxies have been discovered using radio bands (e.g. McGreer et al., 2006; Kuźmicz and Jamrozy, 2021; Delhaize et al., 2021; Lal, 2021). Some radio observations of AGN in closer times have been the result of follow-up projects for already-known objects (Radcliffe et al., 2021). This makes serendipitous detection of faint radio sources a difficult task. With the advent of more powerful instruments and surveys, objects with dimmer radio emission can be detected. But as sensitivity levels are improved, emission from star formation can also be detected, making more difficult the distinction between emission from the AGN and their hosts (Rawlings, 2003).

Recently-developed radio instruments and surveys, such as the Faint Images of the Radio Sky at Twenty-Centimeters (FIRST; Helfand et al., 2015), the EMU pilot survey (PS; Norris et al., 2021), the Karl G. Jansky VLA Sky Survey (VLASS; Lacy et al., 2020; Gordon et al., 2020), and the LOFAR Two-metre Sky Survey (LoTSS-DR1; Shimwell et al., 2019), have allowed detection of larger numbers of RGs (e.g. Singh et al., 2014; Williams et al., 2018; Capetti et al., 2020). But determination of some of their properties –e.g. redshift, spectral indices– might still take very long observation times with high sensitivity detectors in, occasionally, other

wavelengths (An et al., 2020). These difficulties make, effectively, characterisation of RGs a costly endeavour.

Introduce fluxes, luminosities and, specifically, radio luminosities

All these surveys and instruments can deliver their measurements as fluxes. Flux (f) is a quantification of the energy from their origin sources that reaches the observer. The amount of energy received depends strongly on the path the light takes from its source. Any intervening material will have some effect, mostly attenuation and obscuration, in the measurements. In order to estimate the effective amount of energy emitted by the source, different assumptions must be made.

The intial formula to obtain luminosities (L) from fluxes as presented by, for example, Carroll and Ostlie (2017) is:

$$F = \frac{L}{4\pi r^2} \tag{1.1}$$

with r being the distance to the emitting source. This expression does not take any change in the path of the light into account and assumes no interactions of it with intervening materials. Rearranging its terms, the luminosity can be written as

$$L = 4\pi r^2 F\,, \tag{1.2}$$

The first modification to be applied is related to the distance at which the source is from the observer. Thus, a correction for its redshift must be applied. Additionally, a term with the correction for the difference between the waveband of the detector and that of the observed source can be included [K-correction;][](Oke and Sandage, 1968). For the specific case of radio bands, it is possible to assume their continuum spectra can be modelled as a power law in the form $S_\nu \propto \nu^{-\alpha}$ (Sommer et al., 2011), with , the spectral index, typically taking values around $\alpha \sim -0.8$. Then, Eq. 1.2 can be written as:

$$L_\nu = 4\pi D_L^2 S_\nu \frac{\mathcal{K}}{(1+z)}\,, \tag{1.3}$$

in which D_L represents the luminosity distance to the source (Hogg, 1999) and $\mathcal{K}$ stands for the aforementioned K-correction. For the radio emission, the K-correction can take the form $(1+z)^{-\alpha}$ (Condon et al., 2002; Radcliffe et al., 2018; Cochrane et al., 2023) and Eq. 1.3 leads to:

$$L_\nu = \frac{4\pi D_L^2}{(1+z)^{1+\alpha}} S_\nu \,. \tag{1.4}$$

And, given that the radio emission has been assumed to have the shape of a power law, it is possible to use this fact to transform luminosities in different frequencies (Delhaize et al., 2017). If a luminosity is measured at a frequency ν_a, the way to convert it into the luminosity at frequency ν_b, using the spectral index, α, is:

$$L_{\nu_b} = \frac{4\pi D_L^2}{(1+z)^{1+\alpha}} \left(\frac{\nu_b}{\nu_a}\right)^{\alpha} S_{\nu_a} \,. \tag{1.5}$$

Translating radio measurements into one single frequency band can help obtaining more data points to create and study more robust radio luminosity distributions. One way to study such luminosity distributions is through Luminosity functions (LFs), which can provide a measure of the evolution of the density of sources in different time (redshift) and brightness (luminosity) bins (e.g. Salpeter, 1955; Schmidt, 1968; Schechter, 1976; Steidel et al., 1999).

As a way to test the number of sources we observe in different wavelengths, redshifts, and luminosities, simulations have been used to obtain an estimate of the number of AGN available to be observed with specific instruments and sensitivities (Habouzit et al., 2022). Some of these simulations (e.g. Amarantidis et al., 2019; Thomas et al., 2021; Bonaldi et al., 2019) have shown that the distribution of AGN and RG along redshift will lead to the detection of a few hundreds of objects per square degree closer to the end of the EoR with deep observations –e.g. Square Kilometre Array (SKA; Braun et al., 2019), which is projected to have μJy point-source sensitivity levels (Prandoni and Seymour, 2015)–.

These expectations of an statistically significant number of AGN and RG in the high-redshift Universe do not match completely with the most recent compilations (e.g. Inayoshi et al., 2020; Ross and Cross, 2020; Fan et al., 2023), which show that close to 300 have been confirmed to exist at redshifts higher than 6 in the whole sky. This mismatch emphasises the need to detect and confirm the presence of more AGN than can match models and simulations.

1.1 AGN detection methods

The presence of an AGN can be confirmed (or hinted) in several ways depending on the observed wavelengths. Historically, one of the first wavelengths used to confirm the nature of AGN, and the dust enshrouding them, was IR (for a historical review, see Sajina et al., 2022).

Assuming that the activity in SMBHs and some components of their host galaxies are correlated (see, for instance, Magorrian et al., 1998; Ferrarese and Merritt, 2000; Gebhardt et al., 2000; Häring and Rix, 2004; Gültekin et al., 2009; Beifiori et al., 2012; McConnell and Ma, 2013; Kormendy and Ho, 2013; Heckman and Best, 2014, and references therein), and the current unified model for AGN (Urry and Padovani, 1995; Netzer, 2015), most of the activity from the accretion in AGN will be obscured by a dusty torus surrounding the SMBH (e.g. Lacy and Sajina, 2020). The peak of this activity will be correlated with that of the star formation in the host galaxy, thus, increasing the fraction of obscured light observed in such systems (Madau and Dickinson, 2014). In this way, the highest probability of detecting an AGN will be by observing in infrared wavelengths. **Expand on IR observations of AGN.**

As mentioned previously, X-ray is considered as an efficient way to confirm the presence of an AGN (e.g. Andonie et al., 2022). Based upon either the extension or the intensity of their emission, sources can be identified as AGN without large uncertainties (LSST Science Collaboration et al., 2009; Padovani et al., 2017; Maitra et al., 2019; Osorio-Clavijo et al., 2023). If an X-ray point source has a luminosity higher than $\sim 10^{42}\,\mathrm{erg\,s^{-1}}$, it is highly likely to be an AGN, thus several sources have been detected in this way **Add examples of X-ray detections of AGN.**.

Many traditional AGN detection methods make use of spectral or photometric observations of objects which, based upon several criteria, determine their nature or class (Padovani et al., 2017; Hickox and Alexander, 2018; Pouliasis, 2020; Chaves-Montero et al., 2017). In the case of spectroscopy, Optical and IR observations have been used to look for the presence of emission lines that might indicate activity from AGN in their spectra (Magliocchetti, 2022). This method provides the best way to determine the presence of an AGN. One method derived from spectroscopic observations is the use of the Baldwin-Phillips-Terlevich (BPT; Baldwin, Phillips, and Terlevich, 1981) diagram, which has been used extensively to detect and diagnose AGN and the SMBH they host based on detected emission lines (e.g. Toba et al., 2014; Sartori et al., 2015; Latimer et al., 2021; Birchall et al., 2020; Ceccarelli et al., 2022). The BPT diagram uses ratios of optical emission lines [O III] $\lambda5007/\mathrm{H}\beta$, [N II] $\lambda6584/\mathrm{H}\alpha$, [S II] $\lambda\lambda6717,6731/\mathrm{H}\alpha$, and [O I] $\lambda6300/\mathrm{H}\alpha$ to determine the source of ionisation of the studied sources and separate them between star-forming galaxies and AGN. Further studies have used the BPT diagrams but different thresholds to separate star-forming galaxies and AGN (e.g. Kewley et al., 2001; Kauffmann et al., 2003; Kewley et al., 2006).

Additional diagrams have also been developed with the aim of using different combinations

of emission lines. One remarkable example is the WHAN diagram (Cid Fernandes et al., 2010; Cid Fernandes et al., 2011), which uses que equivalent width of Hα and the [N II] λ6584/Hα line ratio for AGN selection.

In the case of photometry measurements, some of these methods involve the classification of sources using colours (i.e. differences in magnitudes) in different wavebands as a starting point. Usually, one method used to confirm the presence of AGN in a sample is using IR or Near Infrared (NIR) colours. The most highly used data comes from photometric observations carried our with the Wide-field Infrared Survey Explorer (*WISE*; Wright et al., 2010) or *Spitzer* (Werner et al., 2004). Several works have used combinations of *WISE* colours to derive main properties of AGN and their host galaxies (e.g. Stern et al., 2012; Mateos et al., 2012; Assef et al., 2013; Toba et al., 2014; Menzel et al., 2016; Jarrett et al., 2017; Assef et al., 2018; Barrows et al., 2021). With observations from *Spitzer*, similar schemes have been devised (e.g. Lacy et al., 2004; Donley et al., 2012). Based on the combination of measurements, different scales have been developed (e.g. Stern et al., 2005; Donley et al., 2012), which have been extensively used (e.g. Lacy et al., 2013; İkiz et al., 2020; Bonato et al., 2021; Lacy et al., 2021). Additional colour criteria have been developed for the latest and future facilities and observations (e.g. Messias et al., 2012; Langeroodi and Hjorth, 2023, for JWST).

Other techniques to determine the presence of AGN are related to the use of spectral energy distribution (SED) fitting, proper motion measurements, variability, and morphology. In the case of SED fitting, it implies comparing the available photometric measurements of an object to a series of model templates (Pacifici et al., 2023). The models have been constructed using different combinations of properties –e.g. age, metallicity, contribution from different constituents, etc.–. Thus, the examined source will be assumed to have the properties from the model which fits the best. If one of the properties included in the selected template is an AGN, then the studied source will be assumed to be an AGN as well.

High quality astrometric measurements (e.g. the *Gaia* mission; Gaia Collaboration et al., 2016) have allowed using proper motions for the detection of AGN. In particular, the use of the extragalactic content (Gaia Collaboration et al., 2023a) of its data release 3 (DR3; Gaia Collaboration et al., 2023b) has allowed to determine which sources have very small proper motions, compatible with the presence of AGN (e.g. Storey-Fisher et al., 2023).

Another way of assessing AGN is through the use of photometric measurements in different epochs that allow one to also determine the variability scales of a source. AGN present continuum aperiodic variability in all their observed wavelengths in timescales from hours to years (Giveon

et al., 1999). There is evidence of correlation between the AGN variability of fluxes in X-ray, UV, optical, and NIR bands (Uttley et al., 2003; Arévalo et al., 2008; Arévalo et al., 2009; Breedt et al., 2009; Breedt et al., 2010; McHardy et al., 2016; Troyer et al., 2016; Buisson et al., 2017; Suganuma et al., 2006; Koshida et al., 2009; Koshida et al., 2014; Lira et al., 2011; Lira et al., 2015). This variability also depends on luminosity, wavelength, redshift, presence of radio or X-ray emission, and existence of broad-line systems (Vanden Berk et al., 2004). For these reasons, if particular variability patterns are found in multi-wavelength observations of a source, it can be classified as an AGN candidate.

When high spatial resolution observations are used, morphology can be a suitable tool to determine the presence of either an AGN or a star-forming galaxy. It has been found that the presence of an AGN, even when not observed directly, can affect the morphological parameters of its host galaxy (Getachew-Woreta et al., 2022).

In the case of radio emission from AGN, its detection can be triggered by studies in different wavelengths which anticipate such measurement, which is then confirmed by direct observations (e.g. Glikman et al., 2023). Nevertheless, the most used method for the discovery of sources in radio bands is using, directly, observations from radio surveys (Padovani, 2016; Padovani, 2017). As with measurements in other wavelengths, it is possible to obtain radio colours (called and defined accordingly, in this context, spectral indices, Lisenfeld and Völk, 2000), which might help determining whether the emission from a detected source is produced by an AGN or not. In the context of radio measurements, spectral indices are defined as the value of the slope a power law fitted to the radio flux would have (e.g. Zajaček et al., 2019). Bright AGN, for which most of their radio emission come from synchrotron processes, show spectral indices that are similar between them. In this way, it is possible to correlate the measured radio emission with the presence of an AGN (Condon, 1992). This might be coupled with studies that show a slight correlation between radio spectral index and radio luminosity for AGN (e.g. Sabater et al., 2019). Special care must be taken with sources that show low levels of radio emission. As explored by, for instance, textcites**missing references**, radio sources with low luminosities can have, as the source of their emission, both the AGN they host or star formation events. Above a threshold in luminosity, it might be said that the radio emission detected in a source has been originated from the AGN. Several thresholds have been proposed using different approaches. Most of them have been devised for low-redshift regimes using the distributions of both AGN and star-forming galaxies (via LFs). When the derived luminosities are close or below a certain threshold, **Mention threshold in radio luminosities for classifying sources as**

radio AGN. the fraction of the emission budget that comes from star-formation events increases. **Such change makes the definition of an AGN candidate harder**.

Usually, the opposite process is also performed. It implies searching for radio detections and, afterwards, classifying them as AGN (or any other kind of source). This procedure is based upon analysing the structure of the studied images and looking for features that might indicate the presence of an AGN (for instance, from their radio jets). Several tools have been developed to attain this goal. For instance, Python Blob Detector and Source Finder (PyBDSF; Mohan and Rafferty, 2015), Blobcat (Hales et al., 2012b; Hales et al., 2012a), and Aegean (Hancock et al., 2012; Hancock et al., 2018). In general, these tools look for islands of emission in images and, depending on the selected detection level, they can merge these structures and create larger objects that can be linked to astrophysical sources.

Explanation of how these tools work

1.2 Redshift determination

In order to determine a precise distribution of AGN across cosmic time, unambiguous redshift measurements are needed (e.g. Naidoo et al., 2023). Spectroscopic redshifts, being the most precise measurements, can be determined for a large range of objects, from supernovae (e.g. Frederiksen et al., 2014; Baltay et al., 2021), to galaxies (e.g. Le Fèvre et al., 2015; Galametz et al., 2013), and AGN (e.g. Rajagopal et al., 2021). Spectroscopic redshifts can be obtained by cross-correlation or fitting of the observed data and set of templates (Tonry and Davis, 1979; Schuecker, 1993; Glazebrook et al., 1998; Aihara et al., 2011; Machado et al., 2013) or by the direct detection and matching of powerful spectral features (Kurtz and Mink, 1998; Stoughton et al., 2002; Garilli et al., 2010). However, their determination can take long and high-quality observations, which are not always available for all sources, rendering them not suited for large-sky catalogues (see, for instance, Silva et al., 2011; Pacifici et al., 2023).

Photometric redshifts are an option which come from the use of photometry measurements and not explicit spectral features of an object (Salvato et al., 2019; Brescia et al., 2021; Newman and Gruen, 2022). In general, they use observations that take less integration time than a comparable spectroscopic measurement and, thus, are used for large surveys that need measurements for large numbers of objects. They are also an option for faint sources.

Photometric redshift methods can deliver a probability for their redshift estimations in the form of a Probability density function (PDF or $p(z)$). These functions can deliver a measure of

the uncertainties that photometric redshifts might have. In general terms, photometric redshifts can be obtained using two different methods: template-based techniques and empirical relations.

Template-based methods come from the fitting of multi-wavelength photometry of a source to a model template (Baum, 1957; Baum, 1962; Loh and Spillar, 1986; Newman and Gruen, 2022; Pacifici et al., 2023). The models have been constructed using different combinations of properties –e.g. age, metallicity, contribution from different constituents, etc.–. Thus, the examined source will be assumed to have the properties from the model which fits the best. However, and depending upon the number and quality of the photometry measurements (e.g. low spectral resolution), these properties can have, sometimes, large uncertainties. Even though this method can use less precise values to determine a redshift, it can take a significative amount of time since it needs to contrast the measured SED to the full set of model templates and, when the number of available measurements is low, the quality of the estimation is largely degraded (e.g. Norris et al., 2019).

Using this method, redshift estimations can be obtained from, for instance, galaxies (e.g. Hernán-Caballero et al., 2021), and AGN (e.g. Ananna et al., 2017; Brescia et al., 2019). As expected, the quality of photometric redshift estimates is highly correlated to the quality of the photometry data used for their determination (Newman et al., 2015; Newman and Gruen, 2022).

Some example tools that use template-based methods to retrieve photometric redshifts are Easy and Accurate $z_{\rm phot}$ from Yale (EAZY; Brammer et al., 2008), Bayesian Photometric Redshifts (BPZ; Benítez, 2000), and Photometric Analysis for Redshift Estimate (Le PHARE; Arnouts et al., 1999; Ilbert et al., 2006).

Examples of SED fitting tools and how they work

For the case of empirical relations, the retrieval of photometric redshifts relies on the use of statistics and large sets of observables (e.g. fluxes and their uncertainties) to determine redshifts and correlations between them which can be used with future observations. Most of these redshift determination methods are related to the use of Machine Learning (ML; Samuel, 1959). These techniques will be further described in this work.

Finally, redshift values can also be determined approximately. Using differences among magnitudes –i.e. colours– it is possible to establish the redshift range in which a source is located. This technique –called drop-out– is, by no means, precise, but can lead to further investigation of sources that are at relevant redshifts ranges for the researcher (with, for instance, the previously described photometric redshift methods). In this way, drop-outs are employed as a mean to generate candidates for pertinent redshift values. Given that it requires no more

calculations than the comparison of some series of colours, it is highly efficient at generating rough redshifts of large samples. It has been, mainly, used 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; Merlin et al., 2021; Uzgil et al., 2021; Champagne et al., 2023; Atek et al., 2023).

Since its first uses, this technique has allowed the detection of high-redshift galaxies (Steidel and Hamilton, 1992; Steidel et al., 1996) through the detection of sharp break in flux between broadband filters that sample the vicinities of the Lyman Break (at a rest-frame wavelength of 912 Å). The location of such break is a function of redshift allowing, thus, to obtain a crude estimate of the redshift for the studied objects.

This page intentionally left blank.

2

Challenges in the analysis of astronomical data

The progress of technology and methods used in Astrophysics has been one of the main drivers for the advancement in our knowledge and understanding of the processes taking place in the Universe. But this undeniable improvement has brought some drawbacks that pose serious challenges that might hinder our ability of retrieving useful results from astronomical data. Most of these problems are rooted in the very large number of new and different observational efforts carried out throughout the years. This abundance of measurements can impact the processes that lead to new calculations and results since more resources and steps are needed to treat a large number of measurements.

As more sources are needed to constrain their properties better, new data sets have been compiled and published. Now, multi-wavelength data are available for large fractions of the sky (e.g Gaia Collaboration et al., 2016; Chambers et al., 2016; Lacy et al., 2020; Kollmeier et al., 2017; Wright et al., 2010; Skrutskie et al., 2006; Abbott et al., 2018). But this profusion of observations has come with new challenges with the most relevant being the volume of data. Lately, analysing all observations one by one has become unfeasible in terms of the time needed to fulfil the task (see, for instance, Brescia et al., 2021). This issue will become greater as future surveys and telescopes are put into service, with relevant examples being the SKA and the Legacy Survey of Space and Time (LSST; LSST Science Collaboration et al., 2009; Ivezić et al., 2019).

Over the last couple of decades, the observational capabilities of single instruments have been improved largely. It has become possible to retrieve measurements of very large areas of the sky without important variations in the observational properties (noise, calibration, etc.). These improvements in the overall properties of observations have made possible the creation of surveys than can cover relevant fractions of the sky. Some examples include the FIRST, the Two Micron All Sky Survey (2MASS; Cutri et al., 2003a; Cutri et al., 2003b; Skrutskie et al., 2006;

Wright et al., 2010), the VLASS, the Pan-STARRS, the Galaxy Evolution Explorer (GALEX; Morrissey et al., 2007), and the AllWISE (AW; Cutri et al., 2013). In the near future, they will be complemented by the LSST in the Vera C. Rubin Observatory, the SKA, the next-generation VLA (ngVLA; Selina et al., 2018; Selina et al., 2023), and *Euclid* (e.g. Euclid Collaboration et al., 2022), and others.

While being able to obtain information from more sources and regions of the sky is, by itself, a very relevant improvement, such number of new measurements to analyse have brought some issues related to the treatment of very large datasets (for a review focused on the challenges of future radio surveys, see Norris, 2017).

2.1 Computational costs

Using very large surveys and catalogues for any sort of calculation involves, accordingly, very high computational costs. Recent observational catalogues might have up to billions of entries with several attributes each and large-area images can cover thousands of square degrees with very high angular resolution (Mickaelian, 2020). Additionally, survey instruments will have data transfer rates well above the normal capabilities of a medium-sized server, reaching up to several TB/day rates (Enke et al., 2012). Dealing with such large datasets requires large amount of resources that are not completely available to everyone (Garofalo et al., 2017).

Additionally, most of the methods traditionally used for the detection, classification, and extraction of properties for astrophysical sources have not been developed to be used with very large catalogues. For this reason, using them in the most recent catalogues and surveys can take restrictive running times that not even the most powerful computing facilities can deal with given their memory or cpu usage. Even if current methods are optimised for their use in large computational facilities, running times would be prohibitively long.

A further factor to consider is that of the energy expense of running such methods for long times in powerful machines. Excesive power consumption can impact negatively, first, in the economical costs of running calculations and, second, in the emission of greenhouse effect gases derived from the energy needed for computation.

Reduction of CO_2 emissions from energy consumption can help alleviating the impact from the climate change (IPCC, 2022). While complete net-zero systems are expected to solve this issue, short-term reduction in the use energy spending are needed to help limiting global warming.

With a focus on ease of use and code readability, `Python` has become a standard language in most of recent astrophysical packages (Astropy Collaboration et al., 2022). Conversely, code written in `Python` tend to be one one of the least efficient in it ecological impact (Portegies Zwart, 2020). Thus, and taking into account that the popularity of `Python` should not decrease in the short-term, it is needed to use techniques and code that can obtain results in shorter times than those available to date.

If needed, use big data citations (Zhang and Zhao, 2015)

2.2 Missing measurements

As with any sort of pysical measurement, a fraction of observations might have issues that can render them unusable for any meaningful calculation (Rubin, 1976; Josse and Reiter, 2018). These problems include malfunction of the detector, incorrect cleaning of the data. If different measurements are to be combined, some sources might have observed in one instance but not in the remaining ones. This might affect the study of time series or multi-wavelength, multi-instrument observations.

Furthermore, some analysis methods require that all measurements need to be available and the lack of one of them can render the full set of quantities from a source useless (Little and Rubin, 2014).

2.3 Multi-wavelength counterpart identification

In the case of observations with different filters or different instruments, a new problem arises. It involves the correct identification of the sources observed in each of the filters. Given that the emission in different wavelengths and different moments in time might come from separate components and processes in the studied objects, each observed instance can present a structure that does not match the others. For this reason, finding and matching counterparts for detected sources can be difficult. This problem is enhanced when observations in several bands and different instruments need to be combined.

Several approaches have been used to mitigate this problem. The simplest of them involves taking the closest source (within a defined search radius) as the counterpart without analysing any physical correlation between the underlying processes that might give rise to the observed emission. If not performed carefully, this technique might introduce identification errors where

overlapping sources can be identified as a single object.

More complex procedures attempt to flag as counterpart the sources with the highest likelihood of being produced in the same position in the space (three dimensional space). One such example corresponds to `NWAY` (Salvato et al., 2018), which can match several catalogues using not only sky coordinates but also additional information such as magnitudes or colours.

3

Machine-assisted pattern detection

Taking into account all the issues that the analysis of large datasets might pose, new tools have been developed as a way to tackle them. In particular, the existence of these major AGN detection, radio measurement, and redshift determination methods raises the need of new techniques which might be able to obtain these properties for large amounts of astrophysical sources with enough precision within a shorter amount of time.

Given that this is a problem suffered by several scientific and, even, non-scientific disciplines (e.g. business-related applications; Costa-Climent et al., 2023), large efforts have been put in order to solve it and many techniques have been developed to deal with the ever-increasing data volumes. New statistical and computer methods can analyse thousands or millions of elements and find relevant trends among their properties (Garofalo et al., 2017). One branch of these techniques is able to, using previously-fed data, predict, with relevant confidence, the behaviour new data will have –i.e. the values of their properties–. This is what has been called ML.

In Astronomy, ML has been used in a wide range of subjects, such as redshift determination (e.g. Nakoneczny et al., 2021; Wenzl et al., 2021), morphological classification (e.g. Ma et al., 2019; Lukic et al., 2019; Mostert et al., 2021; Vardoulaki et al., 2021; Burhanudin et al., 2021), emission prediction (e.g. Dobbels and Baes, 2021), anomaly detection (e.g. Baron and Poznanski, 2017; Giles and Walkowicz, 2019; Lochner and Bassett, 2021; Storey-Fisher et al., 2021; Wagstaff et al., 2022), image reconstruction (e.g. Guglielmetti et al., 2022; Adam et al., 2023; Wilber et al., 2023), observations planning (e.g. Garcia-Piquer et al., 2017; Jia et al., 2023; Sravan et al., 2023), and more (Ball and Brunner, 2010; Baron, 2019; Sen et al., 2022; Huertas-Company and Lanusse, 2023). With ML, it is possible to use previously available measurements and extract useful trends and correlations that can suggest the behaviour of properties from future observations or simulations. ML models are, in general, only fed with measurements and not with physical assumptions (Desai and Strachan, 2021) and they do not need to check the consistency of the predictions or results they provide. This can bring, as a consequence, that running times for this kind of algorithms might be less than typical

physically-based codes (e.g. Buchner, 2019).

3.1 Types of machine-assisted analyses

Concentrating our review on the application of ML, two main branches exist for the application of such techniques. The first of them, called Supervised Learning, deals with the idea that, for each set of measurements, there is a response value that, via modelling, we can predict with some degree of confidence (James et al., 2023). On the other side, Unsupervised Learning refers to the analysis of data that does not have an associated quantity. One of the most popular applications of unsupervised learning is clustering of elements. Modelling data would imply separating them by how similar are their properties (or a combination of them). Then, in the case of supervised learning, further divisions are possible. If the predicted variable (target) is a discrete quantity, this prediction is called a classification. Opposite to that, if the predicted target is continuous, the process is called regression.

In this work, we focus on redshift as one of the regression targets. In this way, most of the further definitions related to regression tasks, will be directly applied to the redshift and its properties.

3.2 Prediction metrics

Several methods exist to assess the results from machine-assisted methods. Most of them, have been designed for supervised tasks. In general, predictions are compared with the original (or true) quantities and a new quantity, a metric, is derived to determine how good the prediction is.

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 (predicted vs. true) matrix where the number of true and predicted classes are compared and results stored in cells with the rate of True Positivess (TPs),

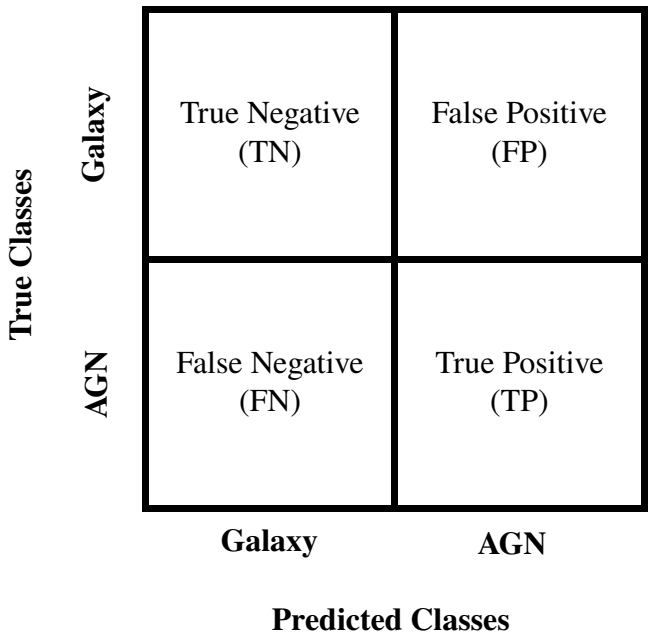


Figure 3.1: Diagram of confusion matrix for the classification of sources between AGN and galaxies.

True Negativess (TNs), False Positivess (FPs), and False Negativess (FNs). An example diagram showing the elements of a confusion matrix is shown ing Table 3.1. An ideal classifier would be represented by a diagonal matrix with no incorrectly predicted elements. As mentioned earlier in Sect. refsec:ML_training, 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 TPs and, consequently, the minimisation of FNs.

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 (i.e. when the fraction of elements of one class is much higher than that of the other) as they take into account the behaviour and correlations among all elements of the confusion matrix. As such, the F1 coefficient is defined as:

$$\mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FN} + \mathrm{FP}}\,. \tag{3.1}$$

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 FNs and FPs. 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:

$$F_\beta = \frac{(1+\beta^2) \times TP}{(1+\beta^2) \times TP + \beta^2 \times FN + FP} \,. \tag{3.2}$$

Using $\beta > 1$, more relevance is given to the optimisation of FNs. When $0 \leq \beta < 1$, the optimisation of FPs is more relevant. If $\beta = 1$, the initial definition of F1 is recovered. As with F1, F_β values can be in the range $[\,0, 1\,]$. Given that we seek to minimise the number of FNs detection, we adopt a conservative value of $\beta = 1.1$, giving more significance to their reduction without removing the aim for FPs. 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:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \,, \tag{3.3}$$

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:

$$\text{Recall} = TPR = \frac{TP}{TP + FN}. \tag{3.4}$$

The True Positive Rate (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):

$$\text{Precision} = \frac{TP}{TP + FP} \,. \tag{3.5}$$

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. 3.2–3.5) are:

$$\mathrm{F}_{\beta}^{\mathrm{no-skill}} = p\,, \tag{3.6}$$

$$\mathrm{MCC}^{\mathrm{no-skill}} = 0\,, \tag{3.7}$$

$$\mathrm{Recall}^{\mathrm{no-skill}} = p\,, \tag{3.8}$$

$$\mathrm{Precision}^{\mathrm{no-skill}} = p\,. \tag{3.9}$$

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

Usually, regression tasks are assessed with the use of metrics such as Mean squared error (MSE), Root mean square error (RMSE), and Mean absolute error (MAE). These measure the deviation of the predicted value from the original quantity. If the original value is called y_{True} and its predicted version is $y_{\mathrm{Predicted}}$, these three regression metrics can be defined as follows.

The MSE is

$$\mathrm{MSE}(y) = \frac{1}{\mathrm{d}} \sum_{i}^{\mathrm{d}} \left(y_{\mathrm{True}} - y_{\mathrm{Predicted}}\right)^2\,, \tag{3.10}$$

with d being the number of elements in the studied sample (i.e. its size). A direct modification of MSE appears when calculating its square root. Then, the root mean squared error is

$$\mathrm{RMSE}(y) = \sqrt{\frac{1}{\mathrm{d}} \sum_{i}^{\mathrm{d}} \left(y_{\mathrm{True}} - y_{\mathrm{Predicted}}\right)^2}\,. \tag{3.11}$$

A third way to quantify the deviation of the predictions from the true values is through the MAE;

$$\mathrm{MAE}(y) = \frac{1}{\mathrm{d}} \sum_{i}^{\mathrm{d}} \left|y_{\mathrm{True}} - y_{\mathrm{Predicted}}\right|\,, \tag{3.12}$$

While MSE and RMSE are sensitive to large deviations in the predictions, the MAE has a linear behaviour with respect to the fluctuations in predicted quantities.

For the case of redshift value determination, the previous metrics are not fully able to

assimilate its logarithmic behaviour. Thus, further modifications are needed in order to use suitable metrics. Namely, a factor must be included to take into account the fact that differences between low redshift values should be penalized more strongly than those at higher redshifts.

We can start with the difference between true ($z_{\rm True}$) and predicted ($z_{\rm Predicted}$) redshift values,

$$\Delta z = z_{\rm True} - z_{\rm Predicted}\,, \tag{3.13}$$

and its normalised difference,

$$\Delta z^{\rm N} = \frac{z_{\rm True} - z_{\rm Predicted}}{1 + z_{\rm True}}\,. \tag{3.14}$$

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:

$$\Delta z_{\rm Total} = \text{median}\,(z_{\rm True} - z_{\rm Predicted}) = \text{median}(\Delta z)\,, \tag{3.15}$$

$$\Delta z^{\rm N}_{\rm Total} = \text{median}\left(\frac{z_{\rm True} - z_{\rm Predicted}}{1 + z_{\rm True}}\right) = \text{median}(\Delta z^{\rm N})\,. \tag{3.16}$$

Using the previous definitions, four additional metrics can be calculated. These are the MAD and Normalised Median Absolute Deviation (NMAD, $\sigma_{\rm NMAD}$; 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, $\sigma_z^{\rm N}$ are typically used. They are defined as:

$$\sigma_{\rm MAD} = 1.48 \times \text{median}\,(|\Delta z|)\,, \tag{3.17}$$

$$\sigma_{\rm NMAD} = 1.48 \times \text{median}\left(\left|\Delta z^{\rm N}\right|\right)\,, \tag{3.18}$$

$$\sigma_z = \sqrt{\frac{1}{\rm d}\sum_i^{\rm d} (\Delta z)^2}\,, \tag{3.19}$$

$$\sigma_z^{\rm N} = \sqrt{\frac{1}{\rm d}\sum_i^{\rm d} \left(\Delta z^{\rm N}\right)^2}\,, \tag{3.20}$$

Additionally, 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 ($\left|\Delta z^{\mathrm{N}}\right|$, Eq. 3.14) 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:

$$\eta = \frac{\#\left(\left|\Delta z^{\mathrm{N}}\right| > 0.15\right)}{\mathrm{d}} \,. \tag{3.21}$$

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

3.3 Classification thresholds

Metrics presented in Sect. 3.2.1 work with a prediction of the status (class) of an element. However, most classifiers deliver a score rather than a definite class prediction. Scores, in the range $[0, 1]$ need to be translated into positive or negative classes. A threshold is defined to separate both states. By default, these models set a threshold at 0.5 in score (which we will call a naive threshold) 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 (Eq. 3.4) given a specific precision limit. By maximising this metric, 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. Thus, a different approach must be taken.

That is possible through the use of the statistical tool called Precision-Recall (PR) curve. Thresholds derived from the PR curves will be labelled as PR. PR curves can help to understand the behaviour of a classifier as a function of its threshold. Both quantities, precision (Eq. 3.5) 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. 3.2). This operation can be performed over the union of training and validation sets, which have been used to create and

train each model.

3.4 Classification calibration

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 (Caruana and Niculescu-Mizil, 2006). 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 (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., 2023).

Calibration of these scores can be done by applying a transformation to their values. For our work, we will apply a Beta transformation. Beta transformation functions have the general form

$$\mu_{beta}(S; a, b, c) = \frac{1}{1 + \frac{1}{\left(e^{c} \frac{S^{a}}{(1-S)^{b}}\right)}}, \tag{3.22}$$

with S being the score from the classifier and a, b, c, free parameters to be optimised. It allows one to re-distribute the scores of a classifier allowing them to get closer to the definition of probability (Kull et al., 2017a; Kull et al., 2017b). 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.

Calibration (or reliability) plots (Niculescu-Mizil and Caruana, 2005) 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 and Smith,

2007; Van Calster et al., 2019).

3.4.1 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:

$$\mathrm{BS} = \frac{1}{\mathrm{d}} \sum_{i}^{\mathrm{d}} (\mathbb{C} - \mathrm{class})^2 \,, \tag{3.23}$$

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 (BS; Glahn and Jorgensen, 1970):

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}} \,. \tag{3.24}$$

In our case, $\mathrm{BS}_{\mathrm{ref}}$ 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 $\mathrm{BSS} \approx 0$, 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 considerably lower than 0, that calibrated transformation will be used as the final scores from the prediction.

3.5 Ensemble learning

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. One technique used to combine the properties of algorithms and improve the results of a prediction is that of ensemble learning. It involves the joint use of individual results from ML models, that have trained to solve the same problem, into one larger model or rule that can deliver a final

prediction (Schapire, 1990; Breiman, 1996; Freund and Schapire, 1996). It has been shown that the combination of several models, and their predictions, can improve the overall prediction results (Opitz and Maclin, 1999).

In order to combine several prediction into a final result, several options are available. The most used, and one of the earliest, way to merge all individual predictions consists of averaging each predicted value into the final prediction (e.g. Sollich and Krogh, 1995). This option is useful for both regression and classification tasks (using its output scores). For the specific case of classification, a voting system can be implemented, where the majority of decisions of the base individual predictors is taken as the final predicted class. This method has been proven to work efficiently (Schapire et al., 1998).

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 (and a third method to combine individual predictions) is called Generalised Stacking (Wolpert, 1992) which can be interpreted as the addition of priors to the model training stage. In this way, two levels of predictors are used. The first level includes all the individual models that are trained on the training set. The second level corresponds to a single model which is trained only on the outputs of the first-level models. Generalised stacking has been applied in several astrophysical problems. That is the case of Zitlau et al. (2016), Cunha and Humphrey (2022), Moya and López-Sastre (2022), Zammit and Adami (2023), Euclid Collaboration et al. (2023a), and Euclid Collaboration et al. (2023b).

3.6 Model explainability and feature importance

Despite the large number of applications it might have, ML has received important criticism related to the lack of interpretability –or explainability, as it is called in ML jargon– of the their 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 (Gerwin, 1974; Langley, 1977; Langley, 1979; Langley et al., 1981; Langley and Zytkow, 1990), which has been developed to extract explicit analytic expressions from the analysed data. Some examples of the use of symbolic regression in astrophysics are

Cranmer et al. (2020), Villaescusa-Navarro et al. (2021), and Cranmer (2023). The lack of explainability 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.

Beside the development of symbolic regression, 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 and Jauhiainen, 2021). It helps to know the relative weights that they have in the decision-making process. In this way, physical insight might be gained about the correlations or triggers between different properties (observed or derived) of the objects of study.

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 subsections.

3.6.1 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; Breiman, 2003). For each feature, the decrease in impurity (a term frequently used in the literature related to ML) 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). 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).

Insight into the decision-making of the pipeline can only rely on the specific weights of the original set of features (see Sect. 6.1).

3.6.2 Local feature importances

As opposed to the global (or 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, for instance, 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, 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 and Tourani, 2020). The combination of Shapley values with several other model explanation methods was used by Lundberg and Lee (2017) to create the SHapley Additive exPlanations (SHAP) values. In this work, SHAP values are calculated using the python package `SHAP`[1] and, in particular, its module for Tree-based predictors (Lundberg et al., 2020). To speed calculations up, the package `FastTreeSHAP`[2] (`v0.1.2`; Yang, 2021) is also used, which allows for multi-thread runs.

One graphical 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.

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; Carvajal et al., 2023a).

[1] https://github.com/slundberg/shap
[2] https://github.com/linkedin/fasttreeshap

4

This thesis

Attending to the difficulties in relating radio emission from AGN with measurements of these sources in additional wavelengths, the main goal of this thesis is to explore and understand possible indicators of the radio emission in AGN from multi-wavelength, multi-instrument, measurements. Thus, we want to develop a process that, in particular, can take information from IR-detected sources (for which there is all-sky coverage with good sensitivity levels) and deliver an indication of whether these sources can correspond to AGN and, more specifically, to radio-detectable AGN.

Given the importance of the detection of AGN in early epochs of the Universe, we also aim to use the aforementioned machinery to derive estimates of photometric redshifts for the sources labelled as prospective radio-detectable AGN. In that way, the focus of our search can be put in the selection of sources as close as possible to the epoch of reionisation. Or at least, in redshift ranges suitable for specific studies.

Having outlined the major issues that exist with the use of new astronomical datasets and surveys (cf. Chapters 1 and 2), the use of machine-assisted techniques (and in particular, machine learning, which aims at the use of available datasets to find relevant trends among their properties to estimate the behaviour of new, unseen data, Samuel, 1959) becomes more relevant than ever before (see Chapter 3). The possibility of analysing very large datasets with reduced computational costs (in time and energy consumption) is one of the main drivers behind the development of this work.

Following the production of candidates for radio-detectable AGN, together with their redshift values, we aim to understand the predictions and how they relate to phyisical properties of the analysed sources. For this goal, we want to apply feature importance analyses to the prediction processes. These techniques can help understanding the inner correlations and trends that allow the creation of several selection rules and prediciton schemes for the creation of predicted sources and some of their properties.

Once the mechanisms leading to the prediction of radio-detectable AGN, and their redshift values, have been understood, we want to apply such indicators and the candidates derived from

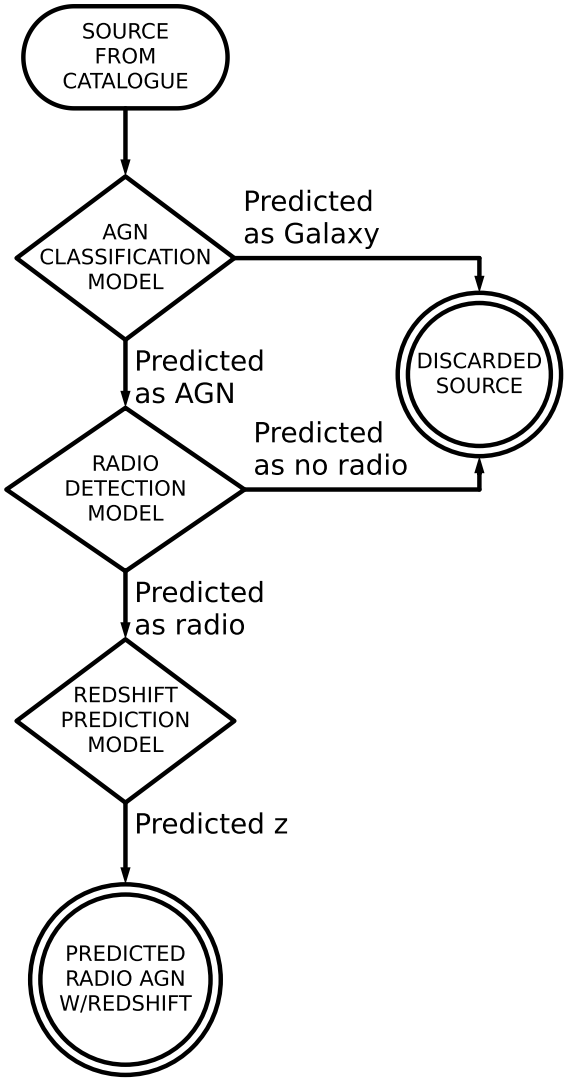


Figure 4.1: Flowchart representing the proposed prediction pipeline used to predict the presence of radio-detected AGN and their redshift values from IR-detected sources.

their use to the analysis and possible solution of different problems related to the observation, classification, and distribution of radio-detected AGN. The application of ML techniques can help creating large collections of candidate radio-AGN that might not have been available previously. The use of such sources might, then, contribute to the improvement of the answers for the questions previously mentioned.

In Fig. 4.1, we present a flowchart of the prediction pipeline we propose for the generation of radio-detectable AGN candidates. We aim to start with a set of IR-detected sources with ancillary multi-band data that can be fed into the a first step that classifies between AGN and galaxies (i.e. not hosting an AGN). Given that we are interested in AGN, we use the predicted AGN and feed them into the second step, which classifies AGN according to their radio detectability. After this step, we select the predicted AGN that have been predicted to be

radio-detectable. A final step uses the predicted radio-detectable AGN and estimates a redshift value for them. Consequently, the prediction pipeline delivers a set of candidate radio-detectable AGN with an redshift estimate.

Given the structure of the pipeline, the ML models in each step will be trained with a different sample of sources. The first step, classification between AGN and galaxies will be trained with all available sources that have been labelled previously as either AGN or galaxies. The second step, classification of radio detection in AGN, will be modelled only with confirmed AGN (with or without radio detections). Finally, the third step of the pipeline, which estimates photometric redhsifts, will be trained with radio-detected AGN. Our focus on radio-detectable AGN is the basis for the omission of the remaining sources in the data sets.

Part of this thesis is based on the work and analyses presented by Carvajal et al. (2021) and Carvajal et al. (2023a). In the following chapters, we present the data sets (as well as the treatment applied to them) used for the generation of the models (Chapter 5), the result of the selection of models and their training (Chapter 6), and the analysis of the use of the prediction pipeline on the selected data sets (Chapter 7). Futhermore, Chapter 8 describes the analysis of the predictions and the models themselves using different approaches. Finally, Chapters 9 and 10 introduce extensions of the use of the results from the prediction pipeline in different contexts as well as future developments to be applied to the prediction pipeline and its individual steps. This thesis concludes with a summary, where final remarks and findings are outlined.

This page intentionally left blank.

5

Modelled datasets

In order to train all models efficiently and test them without fears of obtaining biased metrics, good quality data are needed. This requirement can be translated into selecting a field with coverage in several bands and by diverse instruments. This variety can help the training of the models to cover a broad fraction of the parameter space.

Furthermore, these measurements need to be spread over a sufficiently large area as a way to avoid any bias from using sources that might be connected in some manner. Additionally, in order to validate the predictions from the models, the selected field needs to have a adequate number of sources with certain classifications.

As an extension of the previous requirement, it is desirable that the surveys that cover the selected area also have a much larger footprint. In this way, it will be possible to apply the trained models to different areas of the sky as an further test to our techniques. Thus, all-sky measurements will be prefered over compact (but maybe deeper) surveys.

Finally, and given that one of the goals is predicting radio detectability of sources, the chosen area must have deep and homogeneous radio coverage. If a shallow radio survey is used, the trained model will have access to only a small fraction of possible detections of RGs, which might bias the predictions and their assessment.

5.1 HETDEX Spring field

As training field we selected the area of the Hobby-Eberly Telescope Dark Energy Experiment (HETDEX; Hill et al., 2008) Spring Field covered by the first data release of the LoTSS. The LoTSS survey covers 424 $\deg^2$ in the HETDEX Spring field (hereafter, HETDEX field, see Fig. 5.1) with Low Frequency Array (LOFAR; van Haarlem et al., 2013) 150 MHz observations that have a median sensitivity of 71 μJy/beam. The deep radio observations in this field can help the training stages to retrieve information from a large fraction of sources in the area. HETDEX provides, as well, multi-wavelength homogeneous coverage as described in Sect. 5.3.

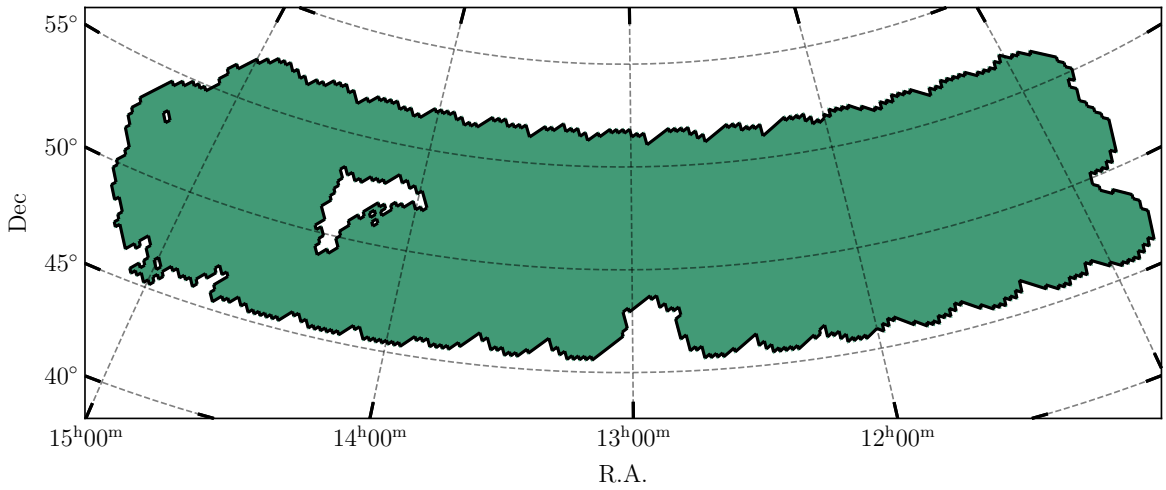


Figure 5.1: Footprint of the area used in the HETDEX field for this work.

5.2 Stripe 82 field

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 SDSS 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 field (see the following section) but with one important caveat: no LoTSS 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. We have selected the S82 field (and, in particular, the area covered by VLAS82, see Fig. 5.2) 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 not exactly the same as those from the training circumstances.

One expected caveat is that, given the shallower nature of the radio observations in S82, the model might predict the radio detection of a source but it might be fainter than the limit from S82. This will be taken into account when comparing metrics between fields.

5.3 Photometry measurements

The base survey from which all the studied sources have been drawn is the CatWISE2020 (CW; Marocco et al., 2021). It lists NIR-detected elements selected from *WISE* and Near-Earth Object WISE (NEOWISE; Mainzer et al., 2011; Mainzer et al., 2014) over the entire

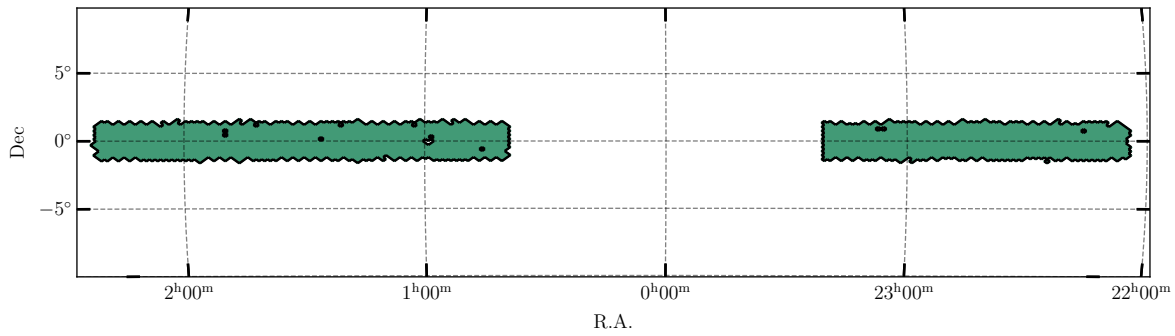


Figure 5.2: Footprint of the area used in the S82 field for this work.

Table 5.1: Bands available for model training in our dataset

Survey	Band (Column name)
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`)

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

sky at 3.4 µm, 4 µm and 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 $\mathrm{mag_{Vega}}$ 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 1.″1 search radius criteria. These catalogues include Pan-STARRS Data Release 1 (PS1; Chambers et al., 2016; Flewelling et al., 2020), 2M, and AW[1]. The adopted search radius corresponds to the distance that has been used by Wright et al. (2010) to match radio sources to PS1 and *WISE* observations and is the smallest Point-Spread Function (PSF) size of the bands included in PS1 (Chambers et al., 2016). Furthermore, the source density of the radio (LOFAR, VLA) and 2M catalogues imply a low statistical (< 1 %) spurious counterpart association, this is not the case for PS1, where the source density is higher. For these reasons, and to mantain a statistically low spurious association between CW and PS1, we limited our search radius to 1.″1. A list of used bands and their origin instruments and surveys is shown in Table 5.1.

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

[1]For the purposes of the analyses, and except when clearly stated otherwise, all photometric measurements were converted to AB magnitudes.

Table 5.2: 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	6380.66	z	10 331.93	H	1335.55	W2	35 700.18
r	9304.58	y	6735.97	Ks	1335.55	W3	14 045.08
i	11 242.35	J	1335.55	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	8249.04	z	13 214.70	H	2330.92	W2	39 025.05
r	12 962.35	y	9226.45	Ks	2330.92	W3	15 393.12
i	14 507.01	J	2330.92	W1	39 025.01	W4	15 472.75

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. 5.5) and no further analyses were performed regarding the detection levels that might be found in any of the fields.

Additionally, we have discarded the measurement errors of all bands. Traditionally, ML algorithms cannot incorporate uncertainties in a straightforward way during training (e.g. Jiang et al., 2021; Michelucci and Venturini, 2023) and, thus, we opted to avoid attempting to use them for training. One significant counter-example corresponds to Gaussian process (GP; Rasmussen and Williams, 2005)s, 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), and 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.

The number of valid measurements in Fig. 5.3 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 (Sects. 5.1 and 5.2). Table 5.2 shows these densities.

5.4 Missing data treatment

In general, ML methods (and their underlying statistical methods, as introduced in Sect. 2.2) cannot work with catalogues that have empty entries (Allison, 2001; Josse et al., 2019). Several techniques have been devised to handle datasets that lack some of their entries. The simplest of them is listwise deletion, which drops all observations that miss, at least, one measurement (Pepinsky, 2018). This process is inneficient given that it reduces the size of the parameter space from which the models can obtain information for its training.

A second method is imputation. Imputation is the process of replacing non-available measurements with substitute values. In general, there is enough information in the remaining entries to derive a meaningful substitute value (Kalton and Kasprzyk, 1982). In this way, typical quantities used for replacement are the mean of the remaining entries or a function of the other meaurements of the same entry.

Depending on the origin of the substitute value, different categories of imputation exist. Using the definitions from Kalton and Kasprzyk (1982) and Chattopadhyay (2017), it is possible to separate it in multiple and single imputation. In turn, single imputation can be divided into mean, random, regression, hot deck, and cold deck imputation.

Single imputation replaces each missing entry with a single value. Opposite to it, multiple imputation, and as initially proposed by Rubin (1987), creates a set of possible values (usually, based upon statisticall arguments) which are included as new instances of the measured object.

Focusing on single imputation, mean imputation replaces all missing measurements with the mean value of the available values. Random imputation uses any value from the existent measurements in the data set to replace the missing entry. Regression imputation uses the full set of available data to derive a possible value for the missing entry. Hot and cold deck imputations replace the missing value with other instances of the data set (or additional data sets in cold deck imputation) that have the same values for the remaining measurements.

Given its simplicity, we have used an ad-hoc variation of single imputation to replace missing values and magnitudes fainter than $5-\sigma$ limits with meaningful quantities that represent the lack of a measurement. We have opted for the inclusion of the same $5-\sigma$ limiting magnitudes as the value to impute with. This method of imputation, with some variations, has been successfully applied and tested, recently, by Arsioli and Dedin (2020), Carvajal et al. (2021), Curran (2022), and Curran et al. (2022).

In this way, observations from 12 non-radio bands were gathered (as listed in Table 5.1).

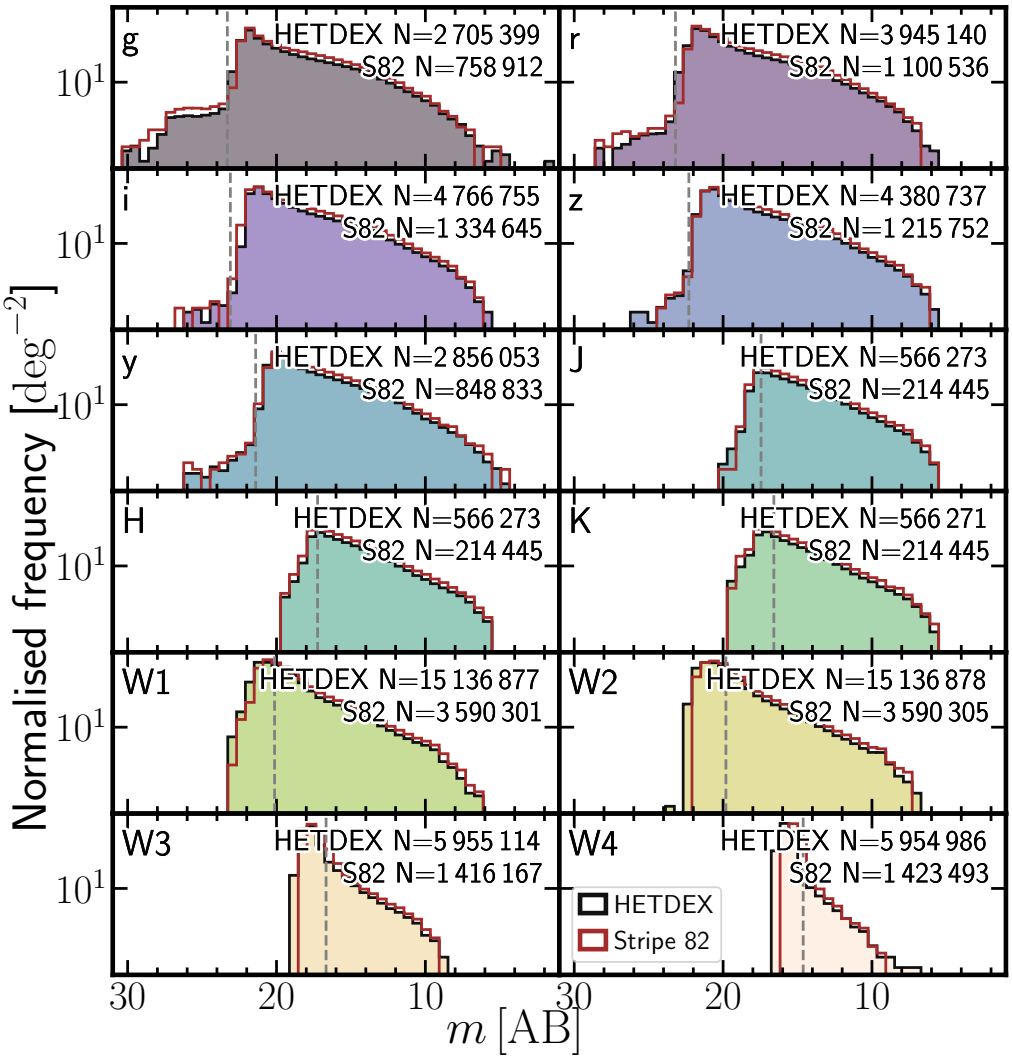


Figure 5.3: Histograms of base collected, non-imputed, non-radio bands for HETDEX (clean, background histograms) and S82 (empty, brown histograms). Each panel shows the distribution of measured magnitudes of detected sources divided by the total area of the field. Dashed, vertical lines represent the $5-\sigma$ 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.

The magnitude density distribution for the sample from the HETDEX and S82 fields, without any imputation, is shown in Fig. 5.3. After imputation, the distribution of magnitudes changes, as shown in Fig. 5.4. Each panel of the figure shows the number of sources which have a measurement above its $5-\sigma$ limit in such band. Additionally, a representation of the observational $5-\sigma$ limits of the bands and surveys used in this work is presented in Fig. 5.5. It is worth noting the depth difference between VLAS82 and LoTSS 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.

Following the same argument of measurement errors, upper limit values have been removed and a missing value is assumed instead.

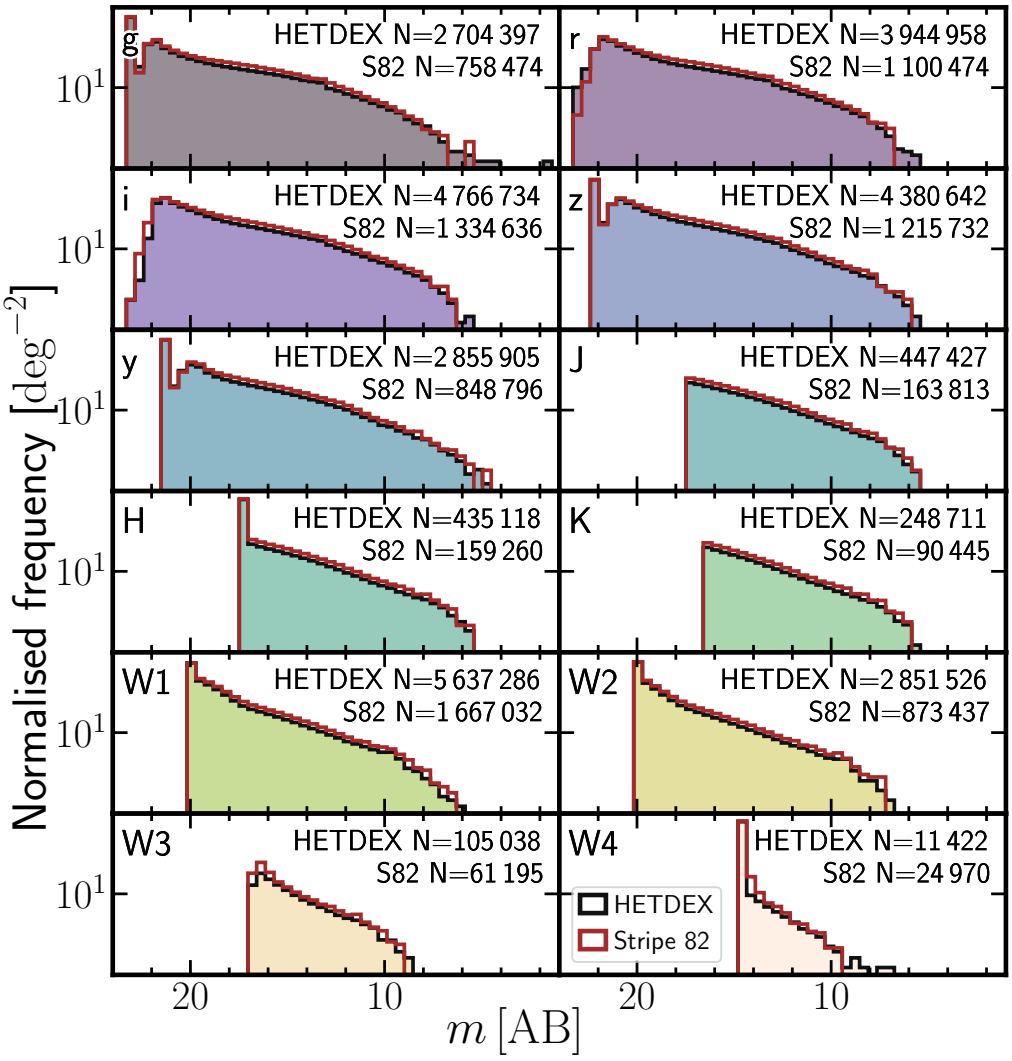


Figure 5.4: Histograms of base collected non-radio bands for HETDEX (clean, background histograms) and S82 (empty, brown histograms) fields. Description as in Fig. 5.3. The number in the upper right corner of each panel shows the number of sources with magnitudes originally measured above the $5-\sigma$ limit included in their corresponding histogram for each field.

5.5 Additional features

As mentioned in the previous sections, each magnitude corresponds to one feature that will be used in the training stages. In order to give the models more information to improve their training, we have generated more quantities as decribed below.

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

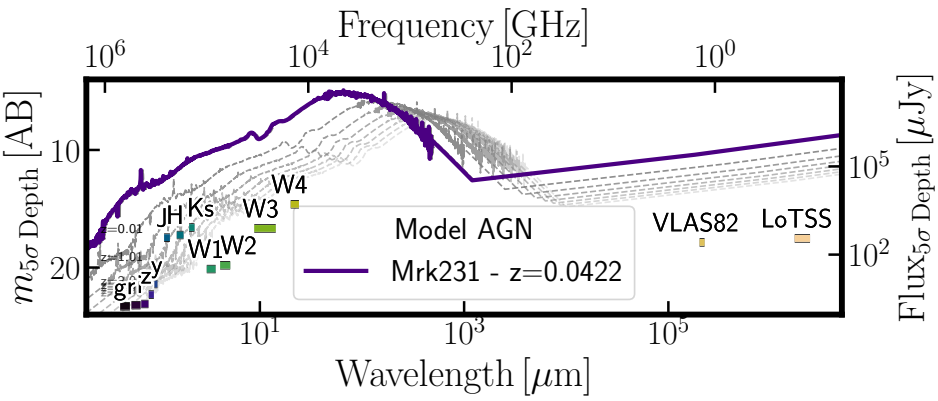


Figure 5.5: Flux and magnitude depths ($5-\sigma$) from the surveys and bands used in this work. Limiting magnitudes and fluxes were obtained from the description of the surveys, as referenced in Sect. 5.3. In purple, rest-frame SED from Mrk231 ($z = 0.0422$, Brown et al., 2019) is displayed as an example AGN. Redshifted (from $z = 0.001$ to $z = 7$) versions of this SED are shown in dashed grey lines.

Table 5.3: Composition of initial catalogue and number of cross matches with additional surveys and catalogues.

Survey	HETDEX	Stripe82
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	...	8747
MQC (AGN)	50 538	17 743
SDSS (Galaxy)	68 196	4085

AGN in HETDEX and 17 743 confirmed AGN in S82.

Similarly, the sources in our parent catalogue were cross-matched with the SDSS 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`) and no magnitude measurements were extracted from it. 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, SDSS-DR16 identifies 4085 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 5.3.

Then, colours from measured and imputed magnitudes were considered were added as

features. We created 66 of them, which correspond to all the available combinations of two magnitudes between the 12 selected bands, symbolised by the expression $\binom{12}{2}$. These colours are labelled in the form `X_Y` where `X` and `Y` are the respective magnitudes. Depending on the stage of the training process, the number of used colour might be reduced.

An additional feature shows the number of non-radio bands in which a source has valid (i.e. non imputed) measurements. We have called it `band_num` and it has been produced counting the number of valid values that each source showed before imputation. The values of this feature can range from 2 (source only detected in CW) up to 12 (source detected in every band of all selected surveys). This feature could be, very loosely, assimilated to the total flux a source can display. A higher `band_num` will imply that such source can be detected in more bands, implying that it has 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 and Casari, 2018; Duboue, 2020; Sánchez-Sáez et al., 2021; Euclid Collaboration et al., 2023b).

In order to test whether or not a source has been detected in any of the radio surveys we have used in Sect. 5.3, we created a feature, called `radio_detect`, which shows a boolean flag. Its value is `True` (`1`) if we have a valid entry (i.e. a detection) in any of the aforementioned radio catalogues. As such, this flag can only tell if a source can be detected with radio observations similar to the deepest survey from our set and cannot give information of the existence, or not, of radio emission in general.

Lastly, we created an additional boolean feature, called `class`, which shows whether a source has been cross-matched with an element of the MQC or with a galaxy in SDSS-DR16. A value of zero (`0`) means that the source has been found in the SDSS-DR16 galaxy sample, a value of one (`1`) implies that the source has been identified by the MQC. Sources that have not been included neither as AGN nor as galaxies (i.e. unknown sources) have not given any value for `class`. It is worth mentioning that a value of `0` in this flag does not mean directly that a source is not an AGN. It only implies that the studied source has not been listed in the MQC as a confirmed AGN.

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 5.4.

Table 5.4: 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				

5.6 Data re-scaling and normalisation

Attending to the intrinsic differences between ML algorithms, not all of them have the same performance when being trained with features that have absolute values spanning a wide range of values (i.e. several orders of magnitude). In particular, linear modelling of data might overrepresent features with larger absolute values when measuring distances between data point. For this reason, it is customary to re-scale the available values to either be contained within the range $[\,0, 1\,]$ or to have similar distributions (e.g. Toth et al., 1993; Sola and Sevilla, 1997). 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 (i.e. standardisation). Additionally, these new values were power-transformed to resemble a Gaussian distribution. This

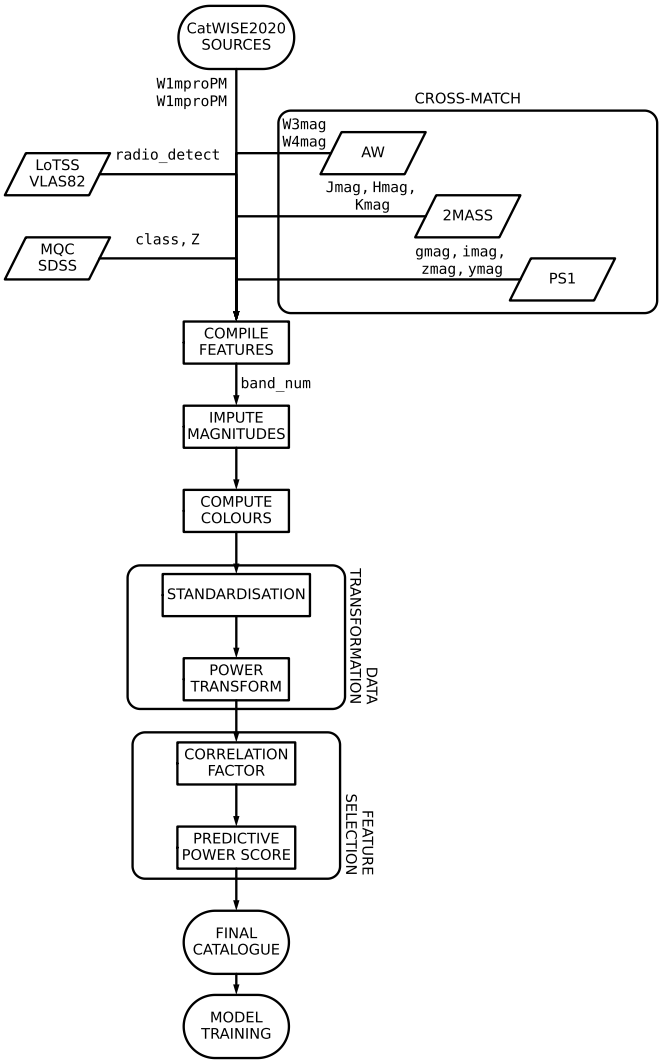


Figure 5.6: Flowchart with steps for data pre-processing. Labels in arrows state the features produced from each step or data catalogue.

transformation helps the models avoid using the distribution of values as additional information for the training. For this work, a Yeo-Johnson transformation (Yeo and Johnson, 2000) was applied.

A representation of the steps performed for the pre-processing of the data (both in HETDEX and S82), and also described in Sects. 5.3, 5.4, and 5.5, is presented in Fig. 5.6.

5.7 Data splitting

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

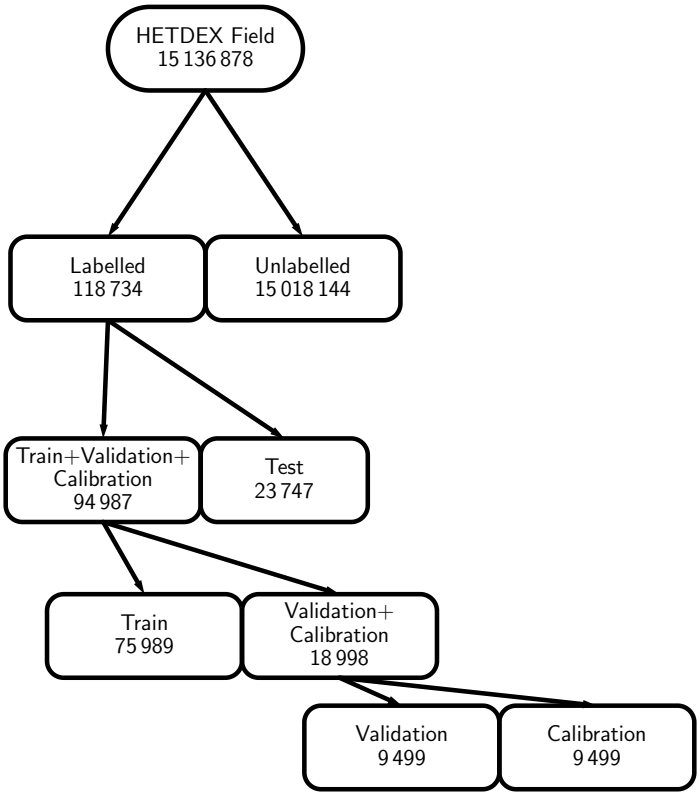


Figure 5.7: Composition of data from HETDEX used for the different steps of this work.

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. Labelled data will be used for training procedures and to, properly, assess the performance of the models.

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). In the case of data from S82, the only separation is done between labelled and unlabelled sources. Given that sources from S82 do not take part into the training of the models, they only need to be isolated from the elements in HETDEX. The splitting process and the number of elements for each subset are shown in Figs. 5.7 and 5.8.

Depending on the model, the needed sources are selected from each of the sub-sets that

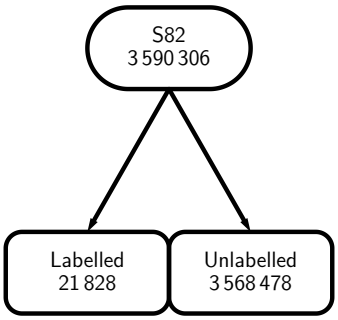


Figure 5.8: Composition of data from S82 used for the different steps of this work.

have been already created. The training set will be used to select algorithms for each step and to optimise their hyper-parameters. 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 and the labelled sources in S82. The use of these subsets will be expanded further in the text.

It is worth noting that the fraction of labelled sources, both in HETDEX and S82, is very low when compared to the total size of the datasets (0.8 % for HETDEX and 0.6 % for S82). These fractions confirm that the problems exhibited in Chapter 2 are ubiquitous and more analysis tools are needed to understand the nature of sources detected in existing catalogues.

This page intentionally left blank.

6

Training of models

In this chapter, we present the results of the creation of the prediction pipeline and the models that make part of it. This procedure includes, for each step, the selection of features, algorithms and hyperparameters.

6.1 Feature selection

ML algorithms, as most data analysis tools, require execution times which increase 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. Feature selection can also help avoiding the inclusion of data that might add noise to the model predictions.

For each model, the process of feature selection begins with 79 base features (Table 5.4) 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.

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. 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). From each pair, we discarded the feature with the lowest Relative standard deviation (RSD; Johnson and 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. Thus, its elimination might not have a large impact the final model.

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. 6.1, 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`.

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`.

Feature selection (cf. Sect. 6.1) was applied to the train+validation subset, with 4612 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`.

6.2 Model stacking

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 and Guestrin, 2016), and Category Boosting (`CatBoost`, `v1.0.5`; Prokhorenkova et al., 2018; Dorogush et al., 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`[1] (`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. Sect. 3.6).

[1] https://scikit-learn.org

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; Friedman, 2002).

6.3 Model training

The procedure described in Sect. 6.2 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 (see last section of Table 5.4). Then, following Michailidis (2017), the hyper-parameters of the stacked models are optimised over the training sub-set (a brief description of this step is presented in Sect. 6.3.1).

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.

The results of model testing for the AGN-galaxy classification are reported in Table 6.1. 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.

Training of the radio detection model was applied only to sources confirmed to be AGN (`class == 1`). The performance of the tested algorithms is shown in Table 6.2. 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 redshift value prediction model was applied to sources confirmed to be radio-detected AGN (i.e. `class == 1` and `radio_detect == 1`). The tested algorithms performed as shown in Table 6.3. 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.

Table 6.1: 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

[a] Metrics obtained using the default probability threshold of 0.5.
[b] Algorithms are sorted by decreasing recall values.
[c] For display purposes, all metrics have been multiplied by 100.
[d] Uncertainties show standard deviation of metrics obtained across all 10 training folds (cf. Sect. 6.2)

Table 6.2: 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

[a] Values and uncertainties as in Table 6.1.

Is is worth noting that, while the use of the mean rank is helpful to select a meta learner, the proper differences in metric values between models are small. These similarities might imply that most algorithms (at least classifiers) can extract the same level of information from the data. For this reason, the use of ensemble learning is justified to help the algorithms extract more information than that they can retrieve on their own.

6.3.1 Hyperparameters optimisation

After the selection of the meta learners of each prediction stage of our pipeline, the predicted values (scores for classifiers and redshift for the regressor) are incorporated to the feature set as new quantities to learn from. Thus, and as shown in Table 5.4, four new feature are added per training instance.

In Table 6.4, 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 optimised using the `SkoptSearch` algorithm embedded in the package `tune-sklearn`[2] (`v0.4.1`; Head et al., 2021), which implements a Bayesian search in the hyper-parameter space.

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

Table 6.3: Results of initial fit for redshift value prediction

Model	$\sigma_{\rm MAD}$ (×100)	$\sigma_{\rm NMAD}$ (×100)	σ_z (×100)	$\sigma_z^{\rm N}$ (×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

[a] Algorithms sorted by increasing $\sigma_{\rm MAD}$ values.
[b] Uncertainties as in Table 6.1.

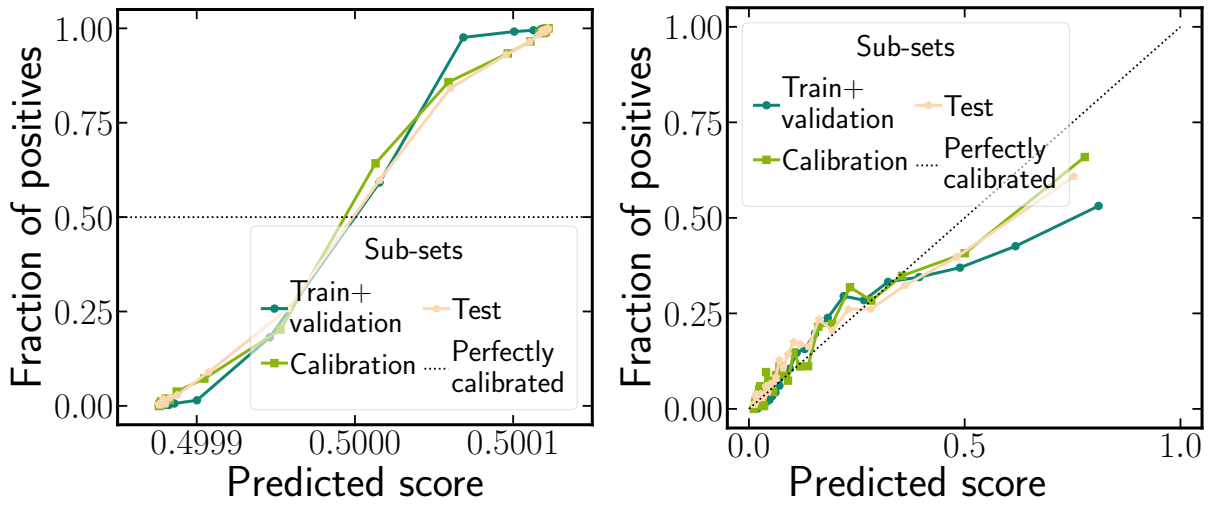
Table 6.4: Hyper-parameters values for meta-learners in modified pipeline after tuning.

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		

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

6.3.2 Calibration of models

In Fig. 6.1, we present the reliability curves for the uncalibrated classifiers. It can be seen that the scores for the AGN-galaxy classifier can be found clustered in a small range around 0.5. This behaviour might indicate some issue with the predictions. But, as presented in Table 6.1, all models used in the training show very high (i.e. satisfactory) metric values. This apparent contradiction might be explained by two factors. First, the sample used for training is highly unbalanced, with most of the sources being labelled as galaxies. Then, there might be a fraction of elements of both classes that share a significant region of the parameter space. When ML algorithms try to classify elements under these two circumstances, they tend to deliver predictions with very low certainties (see, for instance, Vuttipittayamongkol et al., 2021; Santos



(a) AGN-galaxy

(b) Radio detection

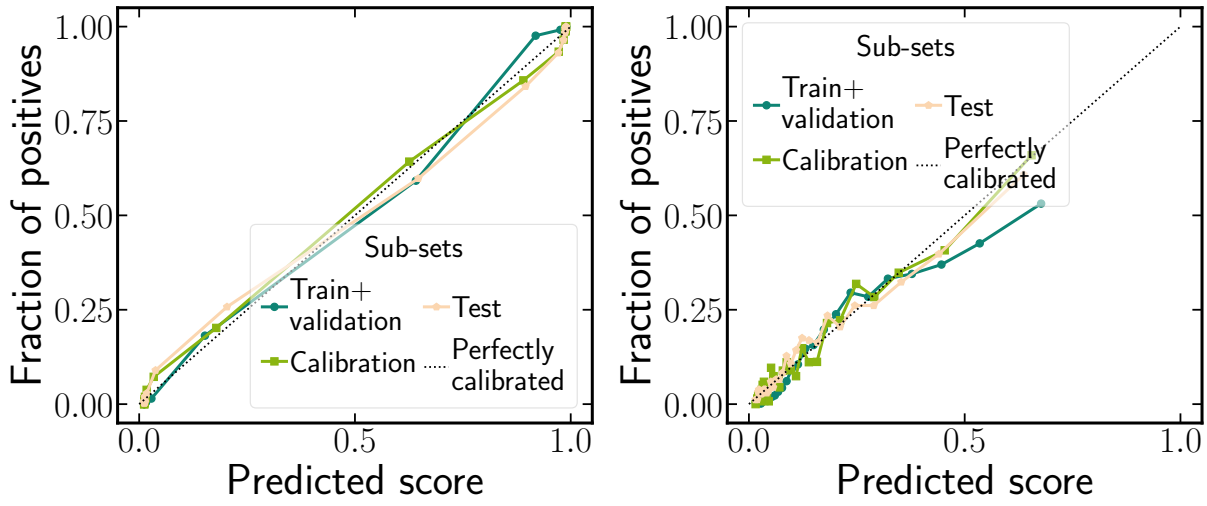
Figure 6.1: Reliability curves for uncalibrated classifiers. Each line represents the calibration curve for each subset in HETDEX field. Data has been binned and each bin (represented by the points) has the same number of elements per curve. Dashed line represents a perfectly calibrated model. (a) AGN-galaxy classification model. (b) Radio detection model.

et al., 2022). This issue can be solved, among other techniques, with the use of probability calibration, which has been implemented in our pipeline.

The previously presented problem does not seem the be present, in the same fashion, in the classification of radio-detectable AGN. There, the distribution of prediction scores ranges from 0.0 up to ~ 0.8. In this case, and given the conditions of the problem of finding indicators of the detection of radio sources from optical and infrared measurements, the source of a lack of scores close to 1.0 can be related to the imposibility of the models of finding stronger connections between all measurements.

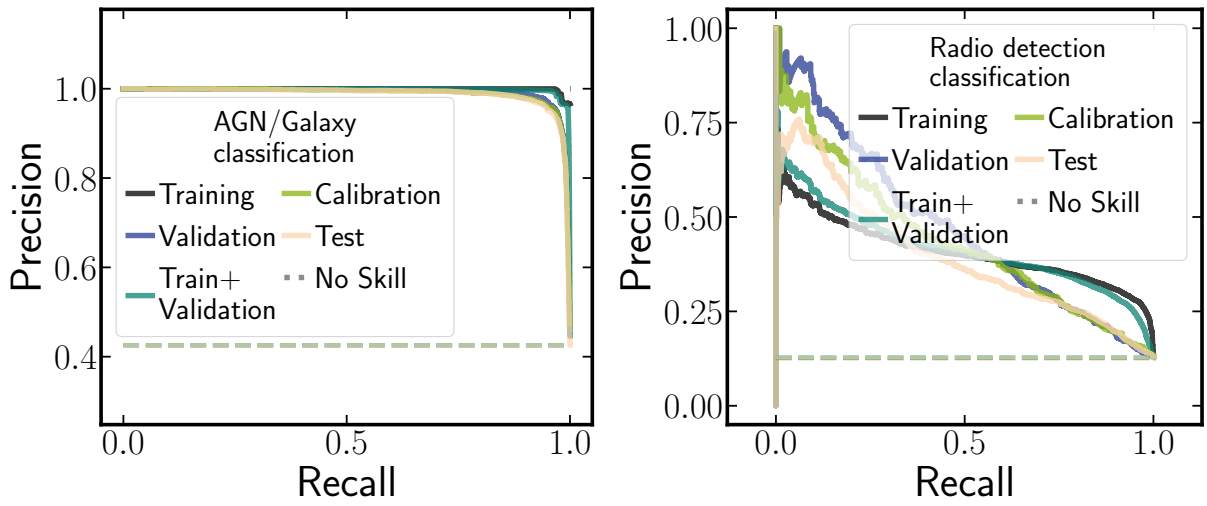
In Fig. 6.2, we present the reliability curves for the calibrated versions of the classifiers. For the AGN-galaxy classifier, the improvement is remarkable. Now, predicted probabilities are distributed in the range $[\,0, 1\,]$ and they follow closely the line of perfect calibration. In the case of the radio detectability prediction, the improvement is milder, with the new probabilities getting closer to the line of perfect probability calibration. Nevertheless, the new probabilities maintain the same distribution as the original scores (between 0.0 and ~ 0.8). This result implies that probability calibration cannot be used to solve issues with the extraction of information from the available features.

From a numerical point of view, 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 intrinsic 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



(a) AGN-galaxy (b) Radio detection

Figure 6.2: Reliability curves for calibrated classifiers. (a) AGN-galaxy classification model. (b) Radio detection model. Details as in Fig. 6.1.



(a) AGN-galaxy (b) Radio detection

Figure 6.3: Precision-Recall curves for the (a) AGN-galaxy and (b) radio detection classification models.

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.

6.3.3 Threshold selection

PR curves for all subsets used in our classification models are shown in Fig. 6.3.

In the case of AGN-galaxy classification, it can be seen that the PR curve does not present any anormality. From the optimisation of the F_β score, the optimal threshold for the calibrated

meta model is 0.348 95. This value was used for the AGN-galaxy model throughout this work.

The PR curves for the calibrated radio-detectability meta model present a diffent behaviour, with noticeable variation among subsets. The optimal threshold for this model is found to be 0.204 60.

7

Prediction of radio-AGN candidates

After training the models, tuning their hyperparameters, and calibrating their scores, we were able to use them for predicting values in data sets which have not been used in the previous stages. In our case, this is the testing sub-set, which is different for each ML problem (classification or regression) and the labelled sources from S82.

7.1 AGN-galaxy classification

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 7.1. The metrics are shown for the use of two different thresholds, the naive value of 0.5 and the PR-derived value of 0.348 95. The confusion matrix (calculated on the testing dataset) is shown in Fig. 7.1.

Overall, the model is able to separate AGN from galaxies with a very high (recall $\gtrsim$94 %) success rate. It is possible to see that the MCC scores for the three sub-sets are in similar levels. That might be an indication of a good training process, in which no substantial over-fitting can be detected.

A closer inspection to the confusion matrix in Fig. 7.1 shows that close to a 45 % of the AGN from the MQC were discarded by our model. And less than 26 % of the predicted AGN are not labelled as such by the MQC. An in-depth analysis of these results is presented in the following sections.

7.2 Radio detection classification

The application of the stacked model for the prediction of the radio detection of the training, testing, and validation sub-set is summarised in Table 7.2. Similarly, the confusion matrix derived from the prediction results over the test sample is shown in Fig. 7.2.

Table 7.1: Resulting metrics of AGN-galaxy classification model for the test subset and the labelled sources in S82 using two different threshold values. HETDEX and S82 pipeline results are described in Sect. 7.4.

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

[a] All metrics have been multiplied by 100.
[b] Uncertainties show standard deviation of metrics obtained across all 10 training folds (cf. Sect. 6.2)

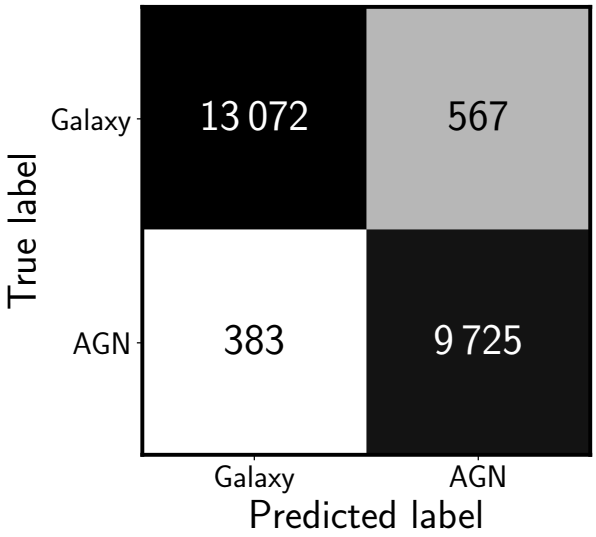


Figure 7.1: Confusion matrix from the results of application of AGN-galaxy classification model to the HETDEX test subset.

7.3 Redshift prediction

In the case of redshift values prediction, the application of the stacked model over the testing sub-set is summarised in Table 7.3. Likewise, the comparison between the original redshift values and those derived from the prediction results is shown in Fig. 7.3.

The results in Table 7.3 show some degree of over-fitting, since the testing scores are a factor of two worse than those from the training and test sub-sets. This happens for all used metrics.

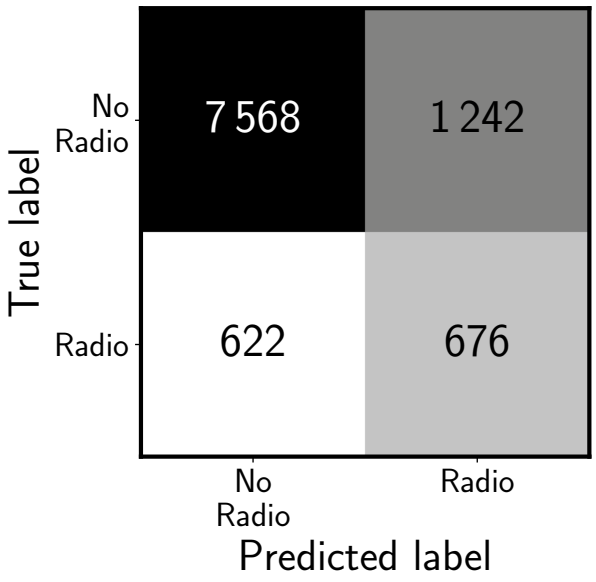


Figure 7.2: Confusion matrix from the results of application of radio-detection classification model for AGN to the HETDEX test subset.

Table 7.2: Resulting metrics of the radio detection model on the test subset and the labelled sources in S82 using two different threshold values. HETDEX and S82 pipeline results shown as part of the discussion in Sect. 7.4

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

[a] Values and uncertainties as in Table 7.1.

7.4 Prediction from pipeline

The sequential combination of the models described in Sect. 6.3 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 Figs. 7.4, 7.5, 7.6, and 7.7 and Tables 7.3 and 7.4. Overall, we observe

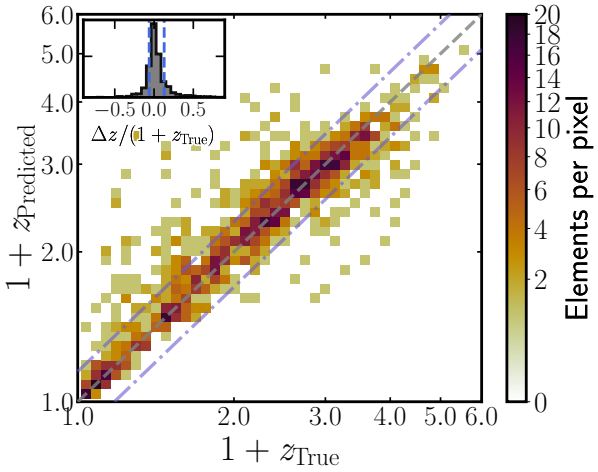


Figure 7.3: Density plot of comparison between original and predicted redshifts from the results of application of redshift prediction model to testing subset. Each point is colour-coded following the colorbar. Grey, dashed line shows $x = y$ relation and purple, dot-dashed lines show the limits where outliers are defined (cf. Eqn. 3.21). Inset shows the distribution of $\Delta z^{\rm N}$ values from the points shown in main plot, with a $\langle \Delta z^{\rm N} \rangle = 0.0442$.

Table 7.3: Redshift prediction metrics for the test subset from HETDEX and S82 labelled sources as discussed in Sect. 7.4

Subset	$\sigma_{\rm MAD}$ (×100)	$\sigma_{\rm NMAD}$ (×100)	σ_z (×100)	$\sigma_z^{\rm N}$ (×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

[a] Values and uncertainties as in Table 7.1.

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, 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 A.2 and A.3 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. 7.8. The pipeline outputs for a small sample of the predicted radio AGN are presented in Tables A.4 and

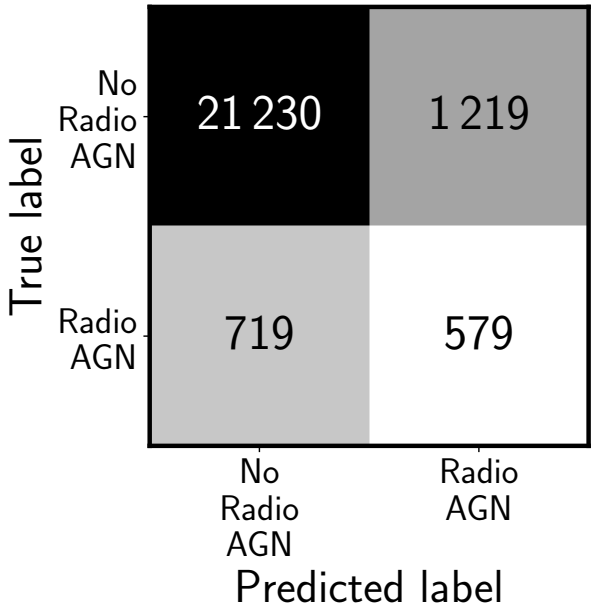


Figure 7.4: Combined confusion matrix from the full radio-AGN detectability prediction computed using the testing subset from HETDEX.

Table 7.4: 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

[a] Values and uncertainties as in Table 7.1.

A.5 for HETDEX and S82 respectively.

7.5 No-skill classification

As presented in Sect. 3.2.1, Eqs. 3.6–3.9 show the base results for a classification with no skill. Table 7.5 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 7.5 coincide with those exhibited in Tables 7.1, 7.2, and 7.4. 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.

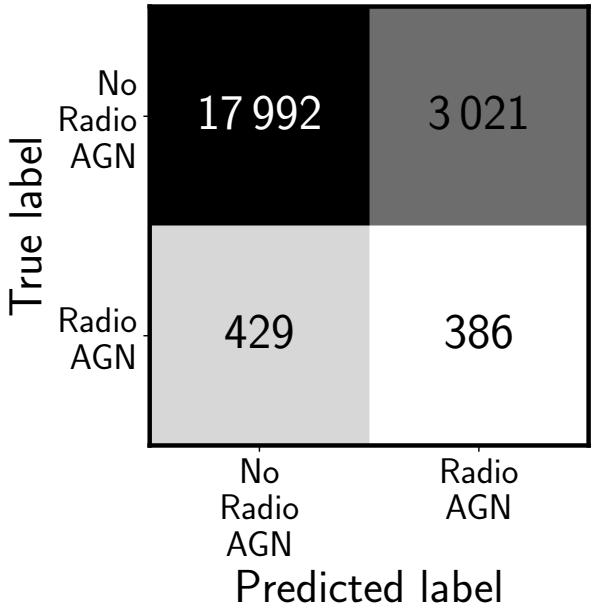


Figure 7.5: Combined confusion matrix from the full radio-AGN detectability prediction computed using the labelled sources from the S82 field.

Table 7.5: 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	5.47	0.00	5.47	5.47
S82	AGN-galaxy	81.29	0.00	81.29	81.29
	Radio-detection (label)	4.59	0.00	4.59	4.59
	Radio AGN	3.73	0.00	3.73	3.73

[a] All metrics have been multiplied by 100.

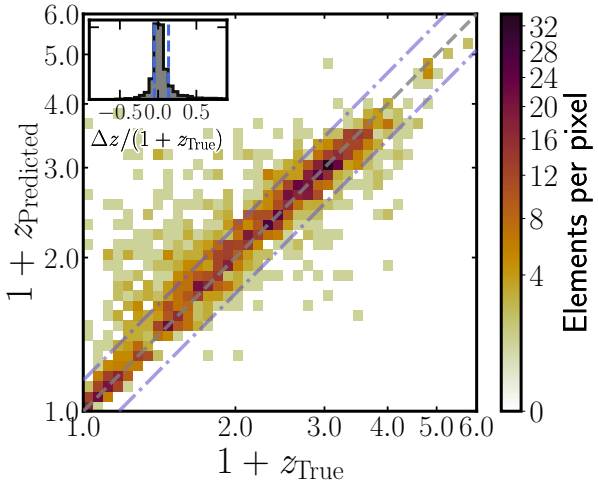


Figure 7.6: Density plot of comparison between original and predicted redshifts from the results of application of redshift prediction model to sources predicted to be radio-detectable AGN in the testing subset. Details as in Fig. 7.3.

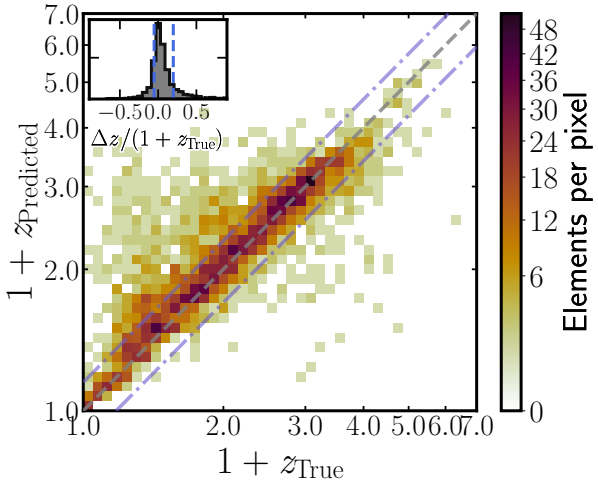


Figure 7.7: Density plot of comparison between original and predicted redshifts from the results of application of redshift prediction model to sources predicted to be radio-detectable AGN among labelled sources in the S82 field. Details as in Fig. 7.3.

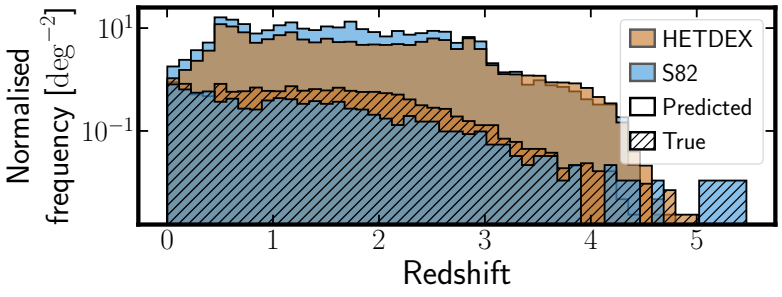


Figure 7.8: Redshift density distribution of the predicted radio-AGN within the unlabelled sources (clean histograms) in HETDEX (ochre histograms) and S82 (blue histograms) and true redshifts from labelled radio-AGN (dashed histograms).

This page intentionally left blank.

8

Analysis of prediction method

8.1 Comparison with previous works

In this section, 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.

8.1.1 AGN detection prediction

In order to understand the significance of our results and ways for future improvement, we separate the comparison with previous works in two parts. First, we present previously published results from traditional methodologies. In second place, we offer a comparison with ML methods.

Traditional AGN selection methods are based on the comparison of the measured SED photometry to a template library (Walcher et al., 2011). A recent example of its application is presented by Thorne et al. (2022b) 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., 2018a) and the Galaxy and Mass Assembly (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 MIR 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 MIR colour selection due to its potential to detect hidden or heavily obscured AGN activity.

Based on *WISE* data, Stern et al. (2012, hereafter S12) proposed a threshold at W1 - W2 $\geq$ 0.8 to separate AGN from non-AGN using data from AGN in the Cosmic Evolution Survey

(COSMOS; Scoville et al., 2007) field. A more stringent criterion was developed by Mateos et al. (2012, hereafter 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, hereafter 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, hereafter B18) analysed the quality of MIR 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 8.1 and a graphical representation of the boundaries they create in their respective parameter spaces is presented in Fig. 8.1.

Table 8.1 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.

Figure 8.1 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. 7.1 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

Table 8.1: Results of application of several AGN detection criteria to our testing subset and the labelled sources from the S82 field.

HETDEX test set				
Method	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

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

[a] Naming codes for the used methods are described in the main text (cf. Sect. 8.1.1)
[b] All metrics have been multiplied by 100.

feature set (e.g. X-ray, radio emission). Sources in FP quadrant, 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 7.1). 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 the results from Clarke et al. (2020), Cunha and 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 in this thesis, 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 Data Release 15 (SDSS-DR15; Aguado et al., 2019) photometric data for almost 3 500 000 sources classified as stars, galaxies or QSO. The application of their

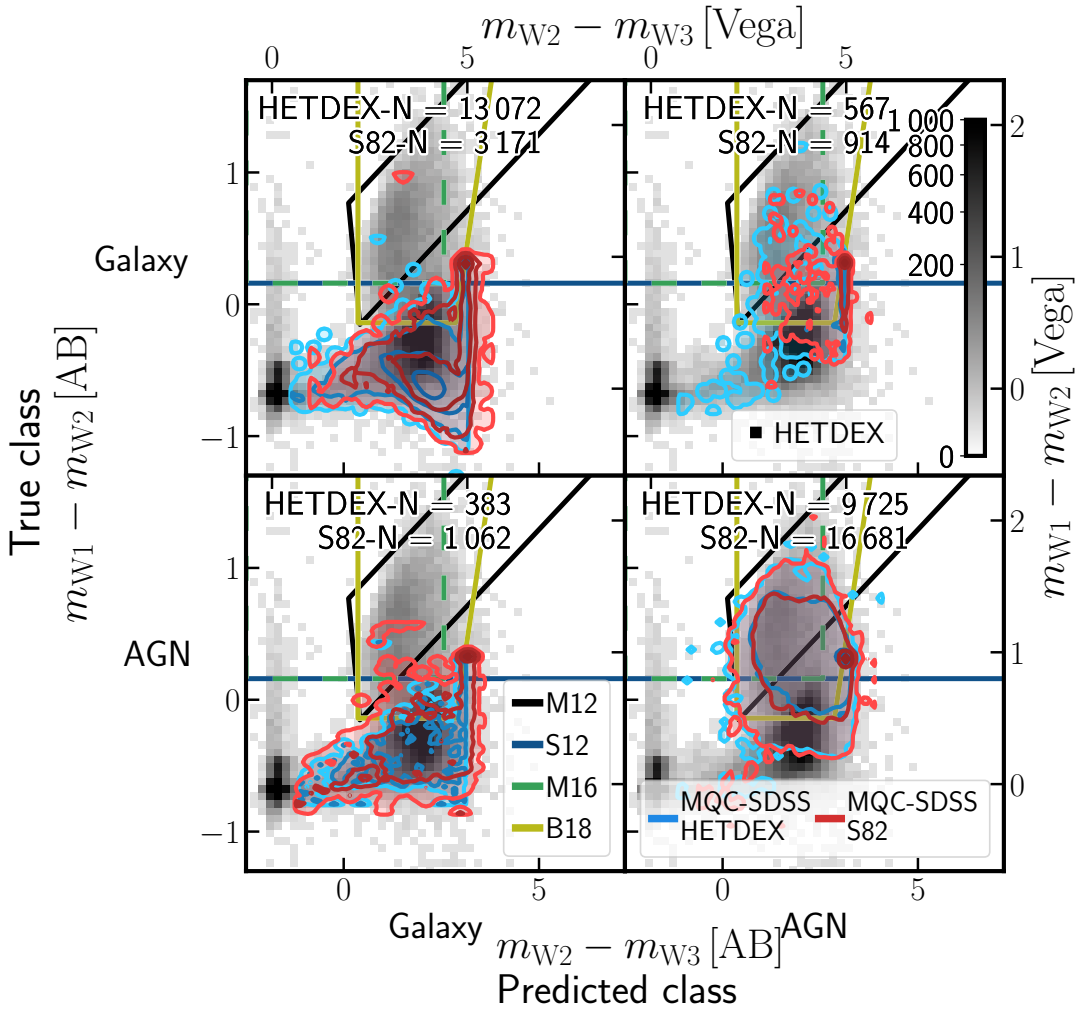


Figure 8.1: W1 - W2, W2 - W3 colour-colour diagrams for sources in the testing subset, from HETDEX, and labelled sources from S82 given their position in the AGN-galaxy confusion matrix (see, for HETDEX, Fig. 7.4 and, for S82, Fig. 7.5). In the background, density plot of all CW-detected sources in the 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. 8.1.1.

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 (right ascension, declination) 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 1500 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 Linear regression (`LR`), Support Vector Machine (`SVM`; Vapnik, 1995; Cortes and Vapnik, 1995), `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.

8.1.2 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 IR emission, Helou et al., 1985; Condon, 1992) and substantial effort has been done towards classifying radio galaxies based upon their morphology (e.g. Aniyan and Thorat, 2017; Wu et al., 2019, Fanaroff-Riley Class I (FRI), Fanaroff-Riley Class II (FRII), bent jets, etc.) and its connection to environment (Miley and 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 $\sim$ 2x success rate of the pipeline to identify radio emission in AGN ($\sim$ 44.61 % recall and $\sim$ 32.20 % precision; see Table 7.4) with the respect to a 'no-skill' or random ($\lessapprox$ 30 %) selection, 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 Sects. 8.3 and 8.4.

8.1.3 Redshift prediction

We compare our results to that of Ananna et al. (2017, Stripe 82X) where the authors analysed multi-wavelength data from more than 6100 X-ray detected AGN from the 31.3 deg^2 of the Stripe 82X survey. They obtained photometric redshifts for almost 6000 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 $\eta = 13.69$ %, values which are similar to our results in HETDEX and S82 except for a better outlier fraction (as shown in Table 7.3, we obtain $\eta_{S82} = 25.18$ %, $\sigma_{\rm NMAD}^{\rm HETDEX} = 0.071$, 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 $\eta = 29.72\,\%$, 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. 5.5). 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 (`KNN`; Cover and Hart, 1967) algorithm, 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 EMU. From the different experiments they run, their resulting NMAD values are in the range $\sigma_{\rm NMAD} = 0.05$ to 0.06, and their outlier fraction can be found between $\eta = 7.35\,\%$ and $\eta = 13.88\,\%$. 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 1300 RGs, they have developed regression methods that show NMAD values between $\sigma_{\rm NMAD} = 0.03$ and $\sigma_{\rm NMAD} = 0.06$, $\sigma_z = 0.10$ to 0.19, and outlier fractions of $\eta = 6.36\,\%$ and $\eta = 12.75\,\%$.

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 $\eta = 7\,\%$ and $\eta = 80\,\%$. 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 the HETDEX field, 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 (Brammer et al., 2008; Salvato et al., 2018; Salvato et al., 2011; Brown et al., 2014) and ML-based methods. In particular, they implemented a GP algorithm (GPz; Almosallam et al., 2016b; Almosallam et al., 2016a), 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 $\sigma_{\mathrm{NMAD}} = 0.102$ and an outlier fraction of $\eta = 26.6\,\%$. 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 and Humphrey (2022) also produced photometric redshift predictions for almost 3 500 000 sources (stars, galaxies, and QSO) as part of their pipeline (see Sect. 8.1.1). They combined three algorithms for their predictions: `XGBoost`, `CatBoost`, and Light Gradient Boosting Machine (`LightBGM`; Ke et al., 2017). This procedure leads to $\sigma_{\mathrm{NMAD}} = 0.018$ and $\eta = 2\,\%$. 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 and Humphrey (2022), they applied an additional post-processing step to the redshift predictions attempting to predict and understand the appearance of catastrophic outliers.

8.2 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. 5.4). In Fig. 8.2, 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. 8.2 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

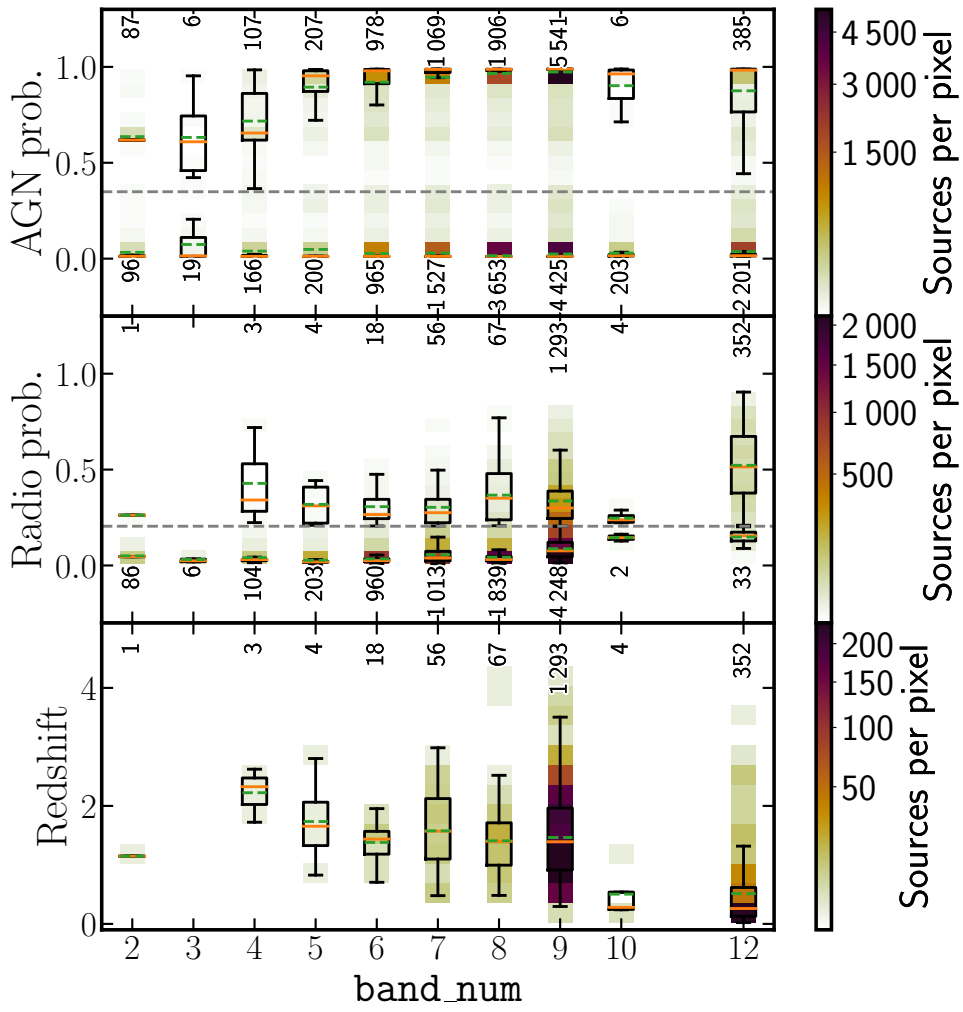


Figure 8.2: Evolution of predicted probabilities (top: probability to be AGN, middle: probability of AGN to be radio-detected) and redshift values for radio-detectable AGN (bottom panel) as function of number of observed bands for sources in test set. In top panel, sources have been divided between those predicted to be AGN and galaxy. In middle panel, sources are divided between predicted AGN that are predicted to be radio-detected 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 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 number of sources considered to create each set of statistics.

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 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.

One interesting feature of all panels in Fig. 8.2 is the lack of sources with `band_num` = 11. This effect can be caused by the inclusion of measurements from 2M. As seen in Fig. 5.3, all three 2M bands have the highest (and almost the same) number of missing measurements. Thus, it is possible to infer that the inclusion of a measurement in one of the 2M bands will imply the addition of the two remaining bands.

In consequence, Fig. 8.2 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. 8.1.3). The AGN-galaxy and radio detection classifications show a mild influence of imputation in their results.

8.3 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; Breiman, 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 ML, for instance, in Breiman, 2001) 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). 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).

Insight into the decision-making of the pipeline can only rely on the specific weight of the original set of features (see Sect. 6.1). Table 8.2 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 8.3) 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, that is, algorithms used and

Table 8.2: 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		

[a] Relative feature importance values are specific to each model training and cannot be compared, numerically, to the values obtained in a different model. A meaningful comparison can be done by contrasting the order in which features are sorted.

stacking procedure. We first notice in Table 8.2 that the order of the features is different for all three models. This difference reinforces the need, as stated in Chapter 4, 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.

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 8.1.1). The *WISE* W3 photometry is though significantly less sensitive than W1, W2 or PS1 (see Fig. 5.5) and a significant number of sources will be represented as upper limits in such plot (see Table 5.3).

One of the main potential uses of the pipeline is its capability to pinpoint radio-detectable

Table 8.3: 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

[a] Relative feature importance values are specific to each model training and cannot be compared, numerically, to the values obtained in a different model. A meaningful comparison can be done by contrasting the order in which features are sorted.

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. 8.3, where sources predicted to be radio-emitting AGN extend to brighter measured W4 magnitudes. This added MIR flux might be simply due to an increased Star Formation Rates (SFRs) in these sources. In fact the 24 µm 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 Star Formation (SF), especially for the sources with real or predicted redshift below $z \sim 1.5$.

Finally, the redshift prediction model shows again that the final estimate is mostly driven by the results of the base learners, accounting for $\sim 82\,\%$ of the predicting power. The overall combined importance of features shows also in this case a strong dependence on several NIR 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 ($z \in [0, 6]$).

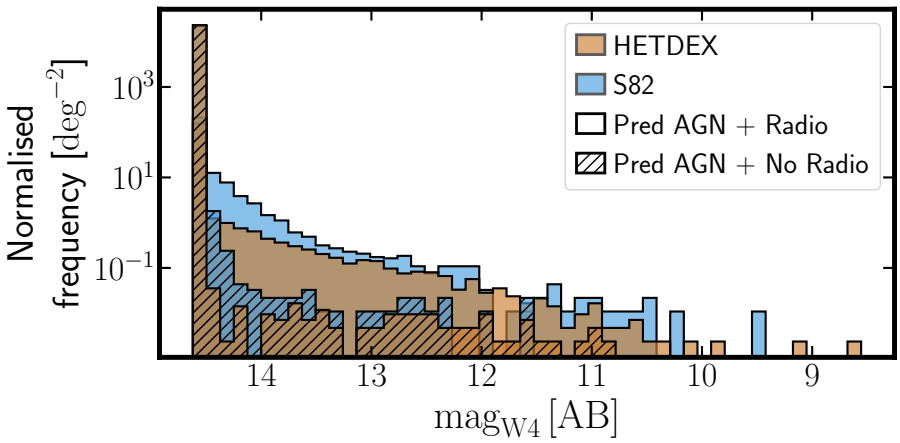


Figure 8.3: W4 magnitudes density distribution of the newly predicted radio-AGN (clean histograms) in HETDEX (ochre histograms) and S82 (blue histograms) and W4 magnitudes from predicted AGN that are predicted to not have radio detection (dashed histograms).

8.4 Local feature importances

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, 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 and Tourani, 2020). The combination of Shapley values with several other model explanation methods was used by Lundberg and Lee (2017) to create the SHAP values. In this work, SHAP values were calculated using the python package `SHAP`[1] and, in particular, its module for Tree-based predictors (Lundberg et al., 2020). To speed calculations up, the package `FastTreeSHAP`[2] (`v0.1.2`; Yang, 2021) was also used, which allows the user to run multi-thread computations.

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. 8.4 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

[1] https://github.com/slundberg/shap
[2] https://github.com/linkedin/fasttreeshap

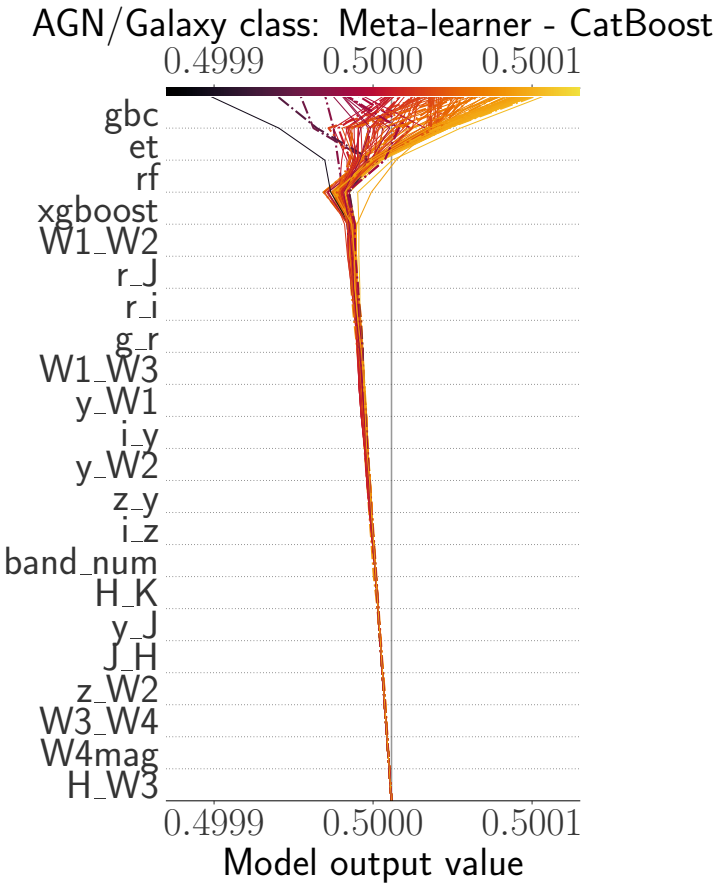


Figure 8.4: Decision plot from SHAP values for AGN-galaxy classification from the 121 high redshift ($z \geq 4$) spectroscopically confirmed AGN in HETDEX. Horizontal axis represents the model's output with a starting value for each source centred on the selected naive threshold for classification. Vertical 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.

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 summarised in Table 8.4. Similarly, SHAP decision plots for the radio-detection and redshift prediction are presented in Figs. 8.5 and 8.6, 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. Figs. 8.7, 8.8, and 8.9 show 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. 8.3, we constructed Table 8.5, 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 8.5 shows, as Table 8.2, that the colour W1 - W2 is the most important discriminator between AGN and galaxies for this specific set of sources. The importance of the

Table 8.4: 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

[a] SHAP values are specific to each model training and cannot be compared, numerically, to the values obtained in a different model and with different data. A meaningful comparison can be done by contrasting the order in which features are sorted.

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. As introduced in Sect. 7.2, radio-detection model shows difficulties when producing a classification based on the provided dataset. This issue is reflected in the narrow decision margin for the non-calibrated stacked model (see model output values –x-axis– close to $\sim$ 0.5 in Fig. 8.5). 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 $M_{\odot}$ 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.

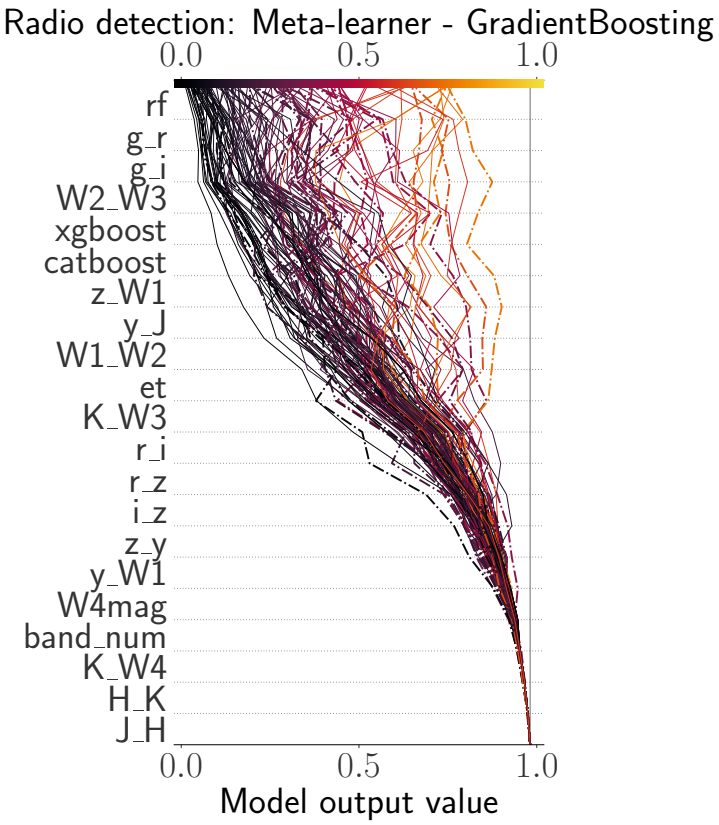


Figure 8.5: Decision plot from the SHAP values for all features from the radio detection model in the 121 high redshift ($z \geq 4$) spectroscopically confirmed AGN from HETDEX. Description as in Fig. 8.4.

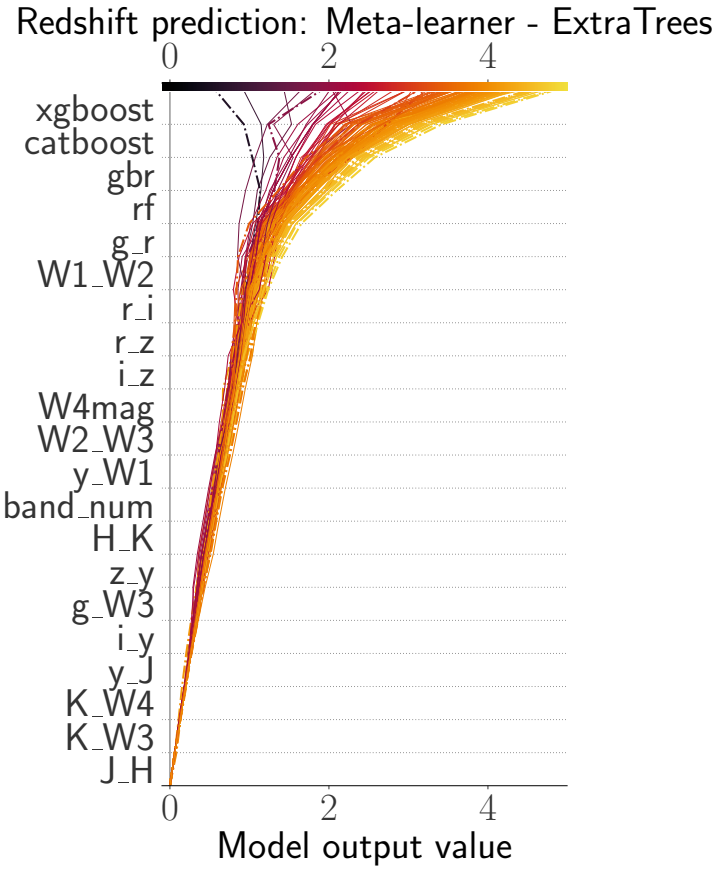


Figure 8.6: Decision plot from the SHAP values for all features from the redshift prediction model in the 121 high redshift ($z \geq 4$) spectroscopically confirmed AGN from HETDEX. Description as in Fig. 8.4.

Table 8.5: 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		

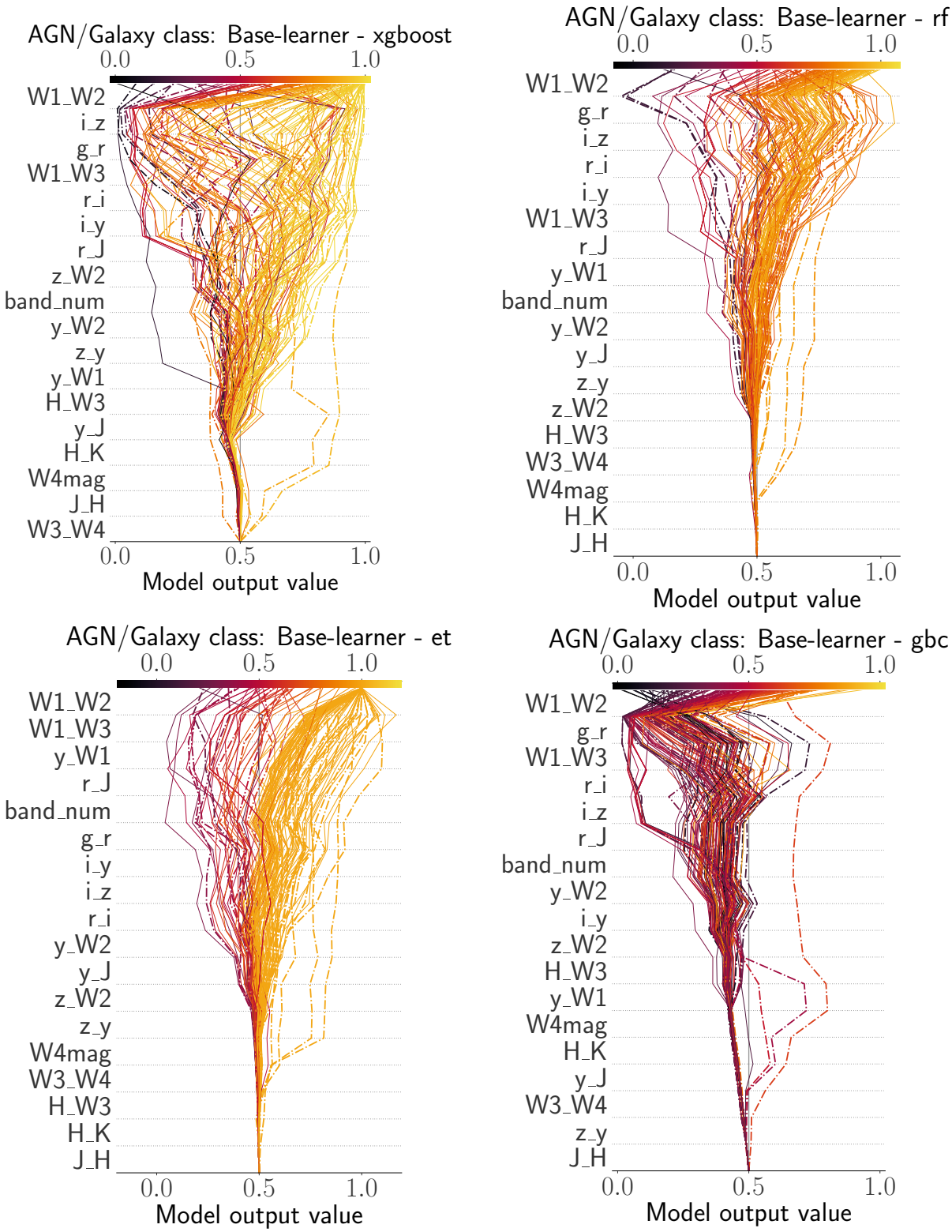


Figure 8.7: SHAP decision plots for base AGN-galaxy algorithms. Details as described in Figs. 8.4. Starting point of predictions is the naive classification threshold. From left to right and from top to bottom, each panel shows the results from XGBoost, RF, ET, and GBC.

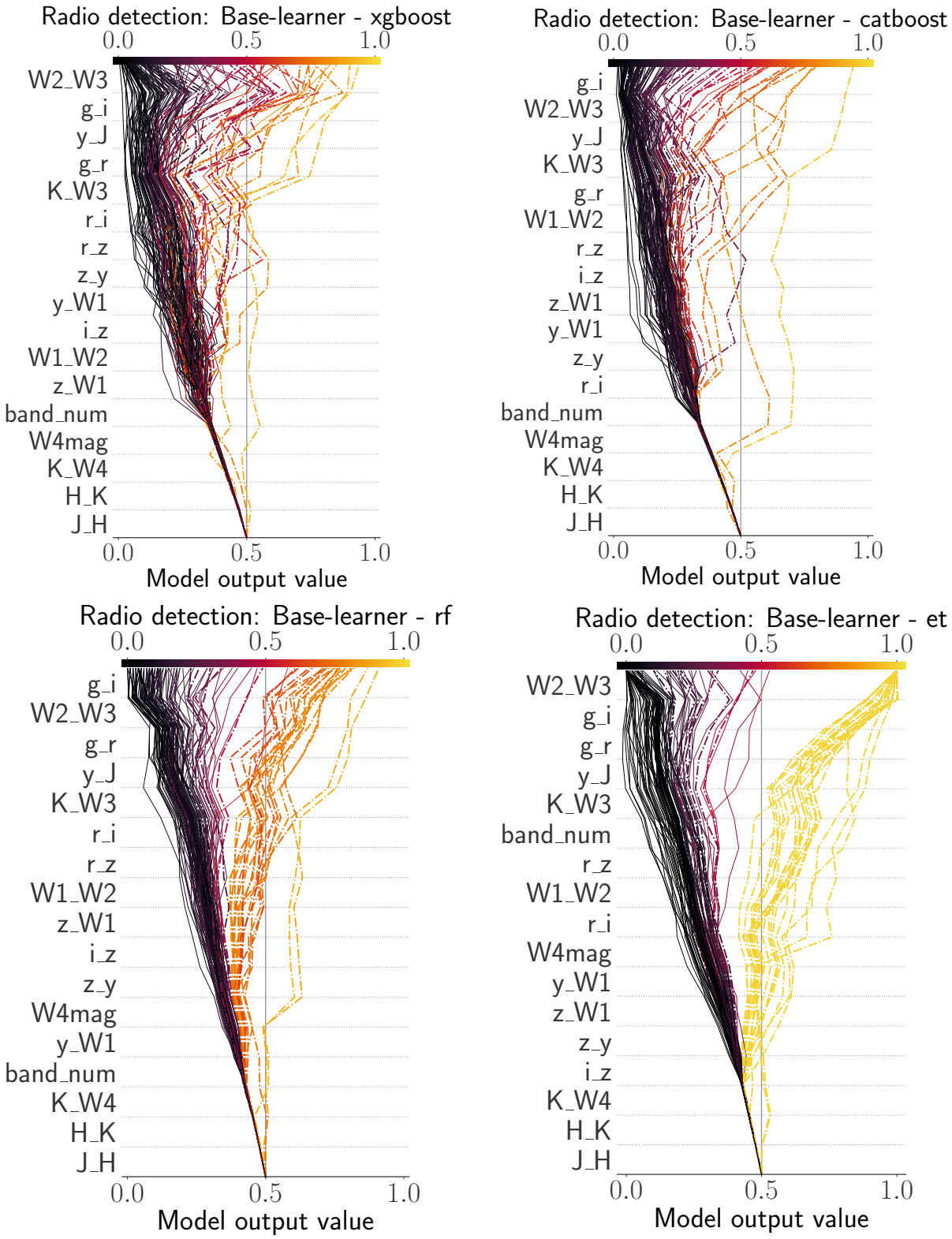


Figure 8.8: SHAP decision plots from base radio algorithms. Details as Figs. 8.4 and 8.7. Each panel with results for XGBoost, CatBoost, RF, and ET.

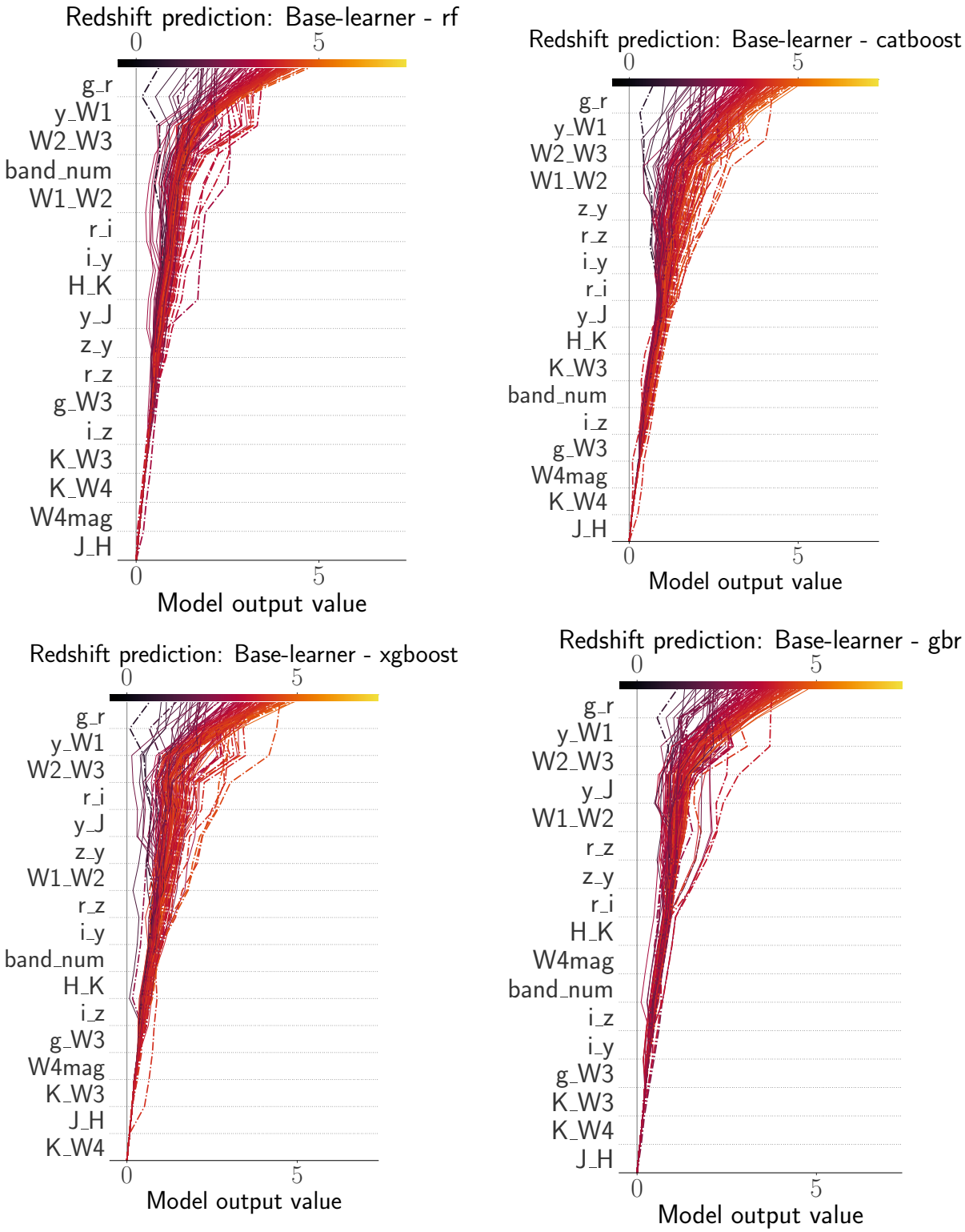


Figure 8.9: SHAP decision plots from base redshift algorithms. Details as in Fig 8.4. Each panel shows results for ET, CatBoost, XGBoost, and GBR.

9

Machine-assisted learning from models

The use of ML methods can be extended beyond the predictions specific models were trained to perform. Given that ML models need to extract correlations and connections between the features they have had access to, they can deliver additional information about the behaviour of the properties of the studied sources (not only those used for training, but also the test or prediction sets). Thanks to its sequential, and fairly independent stages, the prediction pipeline presented in this thesis can be used to extract information for a number of different aspects of the analysis of radio-detectable AGN (together with their redshift values).

In this chapter, we present three instances in which the prediction pipeline can be used to extend the knowledge about our object of study, radio-detectable AGN. In first instance, an example of the extraction of information from the learning process the models use to derive a new, and simple, AGN selection criterion. Then, an example of the use of how the outputs from the prediction pipeline can be directly applied to extract further knowledge from the distribution of radio-detectable AGN in different epochs. And finally, we present the concept of the use of the results of the prediction pipeline for a different goal from its initial purpose, that is, the assessment of multi-wavelength counterparts of radio detections.

9.1 Colour-colour AGN selection criterion

In the introduction of AGN selection methods in Sect. 1.1, it was shown that a combination of photometric colours can be used to determine selection criteria for different sub-sets of AGN. Each one of these criteria is able to retrieve information on specific properties (e.g. redshift) or processes from AGN and their hosts. Additionally, it is possible to extract details from the evolutionary state of some sources by observing the position of them in the colour-colour, or colour-magnitude diagrams.

From the feature importances in Table 8.5 and the values presented in Fig. 5.3 we infer that using optical colours could in principle create novel selection criteria with metrics equivalent to

those shown in Table 8.1 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, cf. Fig. 5.3 and Table 5.2). We tested this hypothesis and derived a selection criterion in the g - r vs W1 - W2 colour-colour parameter space as shown in Fig. 9.1 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 8.1. Their limits are defined by the following expressions:

$$g - r > -0.76\,, \tag{9.1}$$

$$g - r < 1.8\,, \tag{9.2}$$

$$W1 - W2 > 0.227 \times (g - r) + 0.43\,, \tag{9.3}$$

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. These boundaries were drawn to contain, at least, 86 % and, and most 99 % of the AGN (as presented by their distribution contours).

Table 9.1 displays metrics for our new criterion in comparison with previous W1 - W2 vs W2 - W3 AGN colour-colour selection criteria (as presented in Sect. 8.1.1 and Table 8.1). Our classification method 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.

Depending on the degree of available information, the step of the pipeline and the research goals, other combination of colours and features could be used to create more selection criteria for AGN, radio-detected sources, and sources at specific redshift ranges. The use of such methods can be seen as an alternative to the recurrent use of the full set of ML models in our pipeline. Additionally, the creation of these colour-based criteria can be thought as a direct application of the knowledge gathered by the models, rather than a blind application of the latter.

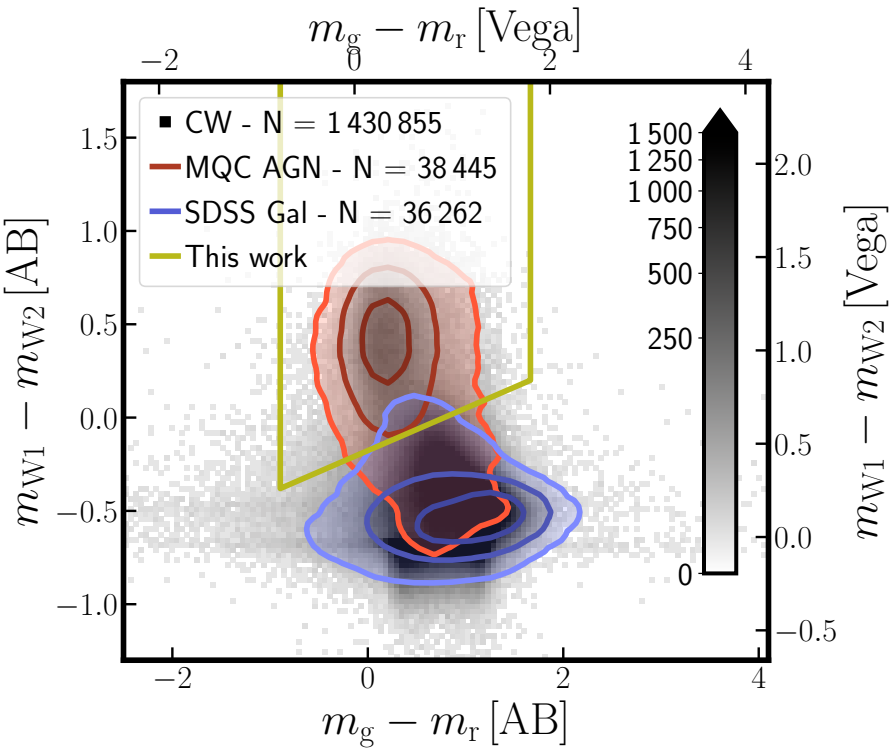


Figure 9.1: AGN classification colour-colour plot in the HETDEX field using CW (W1, W2) and PS1 (g, r) passbands. Grey-scale density plot include all CW detected and non-imputed sources. Red contours highlight the density distribution of the AGN in the MQC and blue contours show the density distribution for the galaxies from SDSS-DR16. Contours are located at 1, 2 and 3σ levels.

9.2 AGN radio luminosity function

A full study of the evolution of the distribution of sources in a region of the sky can be done with the use of LFs, which, using the definition from Salpeter (1955), correspond to the density of sources of a defined class in a range of luminosities (or, equivalently, magnitudes).

Add more context here!

As the performance and inner works of our prediction pipeline have already been assessed (cf. Chapters 6, 7, and 8), we can apply small variations to it as a way to emphasise the goals of this section. For this reason, we decided to run the full training sequence one more time. This time, two changes has been introduced. The first change relates to the cross-match of the radio information for the training set. Instead of using a 1.″1 search radius to find radio detections around the CW-selected sources, we adopted a 6″ circle. With this change, we ensure the selection of radio detections for a larger fraction of sources. This modification is done at the expense of possible miss-identifications of radio counterparts. The second change is the inclusion of a complete new branch of predictions. Taking the description of the pipeline of Fig. 4.1, new models have been added to analyse the sources that have been predicted as galaxies

Table 9.1: Results of application of several AGN detection criteria to our testing subset and the labelled sources from the S82 field. Same as Table 8.1 but including colour-colour criteria from this work.

HETDEX test set				
Method	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

[a] Naming codes for the used methods are described in the main text (cf. Sect. 8.1.1)
[b] Last row of each sub-table corresponds to the criterion derived in this work.
[c] All metrics have been multiplied by 100.

(i.e. not as AGN) by the first model. Thus, these sources will be subject to a prediction of their radio detectability and, those predicted as being radio detectable, will have their photometric redshift values predicted. The inclusion of a new branch is related to the need to compare the radio luminosity distributions between AGN and regular star-forming galaxies across time. A detailed description of the modified datasets and the models produced with them can be found in Appendix B.

As a way to study the distribution of luminosities of AGN as derived from our prediction pipeline, and as part of the efforts to study the behaviour in areas that will be subject to future radio surveys, we have selected, as test field, the area of the EMU-PS. The EMU-PS catalogue[1] has radio information, at 944 MHz from 178 921 compact sources in an area of 270 deg^2 in the southern sky with a depth of 25 μJy/beam to 30 μJy/beam rms and a spatial resolution of 18″ (see Fig. 9.2 for a footprint of the area of EMU-PS).

For the purposes of this exercise, and as presented in Sect. 5.3, we have collected measurements in the EMU-PS area to apply the prediction pipeline. Thus, we have started with the selection of CW-detected sources in the area, finding 10 355 457 detections and successive

[1] EMU-PS data can be obtained from https://doi.org/10.25919/exq5-t894

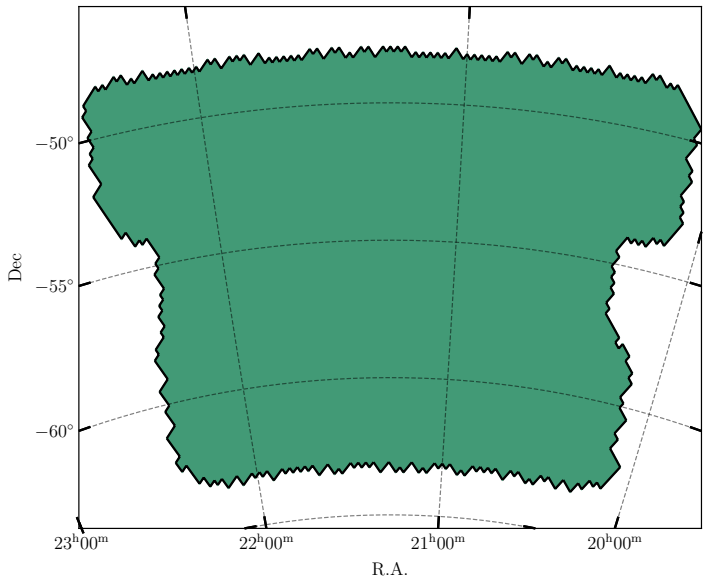


Figure 9.2: Footprint of the area of the EMU-PS field.

cross-matches were performed using a search radius of 1.″1. One exception to this distance was used for the cross-match with the EMU-PS catalogue itself, where a 10″ radius was used instead, which is the maximum search radius used by Norris et al. (2021) to find CW counterparts for their radio detections. The reason for this change is twofold: a larger distance is similar to the size of the restoring beam (18″) and the need for testing the effectiveness of the use of our models for the assessment of radio counterparts. As aforementioned, and additional branch of two models has been included in the prediction pipeline. One that can predict the radio detectability of galaxies (i.e. not AGN) and a second model that can predict photometric redshift values for radio-detected galaxies.

In contrast to the identification of sources in HETDEX and S82 fields and due to their different positions in the sky, the EMU-PS catalogue was cross-matched with alternative catalogues for the association of known AGN and galaxies. In the case of AGN, a more recent version of MQC (v8; Flesch, 2023) was used[2] as well as QSO identifications from the DES Data Release 2 (DES-DR2; Yang and Shen, 2023) and the *Gaia*-unWISE Spectroscopic Quasar catalog (Quaia G20.5; Storey-Fisher et al., 2023) based upon observations from *Gaia* data release 3 (DR3) extragalactic content (Gaia Collaboration et al., 2023a) and the *unWISE* reprocessing (Lang, 2014; Meisner et al., 2019) of the *WISE* data. For the galaxies present in the EMU-PS field, and from the lack of SDSS measurements, we have included the identifications from the final VEXAS Data Release 2 (VEXAS-DR2; Khramtsov et al., 2021).

[2]For this exercise, all sources identified as AGN by MQC were selected, not only those with a redshift measurement, as done in the main model training.

Table 9.2: Composition of initial catalogue and number of cross matches with additional surveys and catalogues in the area of the EMU-PS

Survey	EMU-PS area
CatWISE2020	10 355 457
AllWISE	4 066 594
2MASS	932 926
EMU-PS (10″)	170 702
MQC v8 (AGN)	614
Quaia G20.5 (AGN)	12 491
DES DR2 (AGN)	51 584
VEXAS Spec V2 (Galaxy)	1967

In order to retrieve a large sample of redshift measurements, we included those provided by MQC v8, Quaia G20.5, and spectroscopic redshifts from the Dark Energy Spectroscopic Instrument (DESI) imaging surveys (Dey et al., 2019; Zou et al., 2019), which are contained in the full EMU-PS catalogue. A summary of the number of sources and counterparts found in all different catalogues and surveys can be seen in Table 9.2.

The application of the modified prediction pipeline to the data from the EMU-PS area creates 89 040 candidates to be radio-detectable AGN and 116 400 to be radio-detectable galaxies. Out of them, 36 466 have a radio counterpart in the EMU-PS catalogue (15 123 AGN and 21 343 galaxies), that is, a measured radio flux. To have a sense of the number of predicted sources, the area of the EMU-PS catalogue has 1509 radio-detected AGN and 544 radio-detected galaxies with confirmed identification. These numbers imply a 1000 % and 3900 % increment of sources for radio-detected AGN and radio-detected galaxies respectively. The distribution of predicted photometric redshifts of both samples is depicted in Fig 9.3.

It can be seen in Fig 9.3 that the distribution of radio-detectable galaxies is concentrated between redshift 0 to 1.2. This behaviour corresponds to the original distribution of galaxies used for training in the HETDEX field and thus, to the parameter space coverage of such sources. As expected the model trained with them can only associated new sources to the values in that region of the space of parameters. For the same reason, the distribution of predicted redshifts for radio-detectable AGN spans a larger range, similar to the values of the training in the HETDEX field.

Radio luminosities can be obtained from the use of radio fluxes and redshift values of the sources. Fluxes are obtained from the EMU-PS catalogue, which lists them in its column `flux_int`. A distribution of these values is presented in Fig 9.4. Using Eq. 1.4, they can

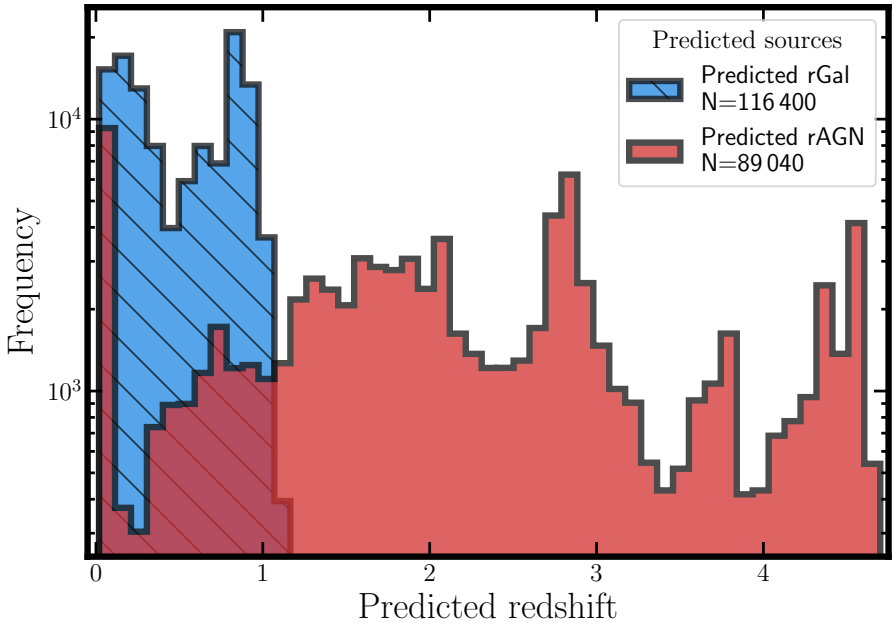


Figure 9.3: Distribution of predicted photometric redshift values for, in hatched blue, predicted radio-detectable galaxies and, in red, predicted radio-detectable AGN in the area of the EMU-PS catalogue.

be transformed into luminosities, at 944 MHz, with the predicted photometric redshift and, either assuming a radio spectral index of $\alpha = -0.7$ (see, for instance, Simpson et al., 2012; Magliocchetti et al., 2014; Šlaus et al., 2020; Mandal et al., 2021; van der Vlugt et al., 2022; Lyu et al., 2022) or using, in the case of EMU-PS, the α values provided by the catalogue. These spectral indices have been calculated from the use of the Taylor terms at the peak pixel of each fitted component (Norris et al., 2021). In order to obtain the luminosity distances for the sources, we have adopted a flat Λ cold dark matter (ΛCDM) cosmology, with $\Omega_m = 0.31$, $\Omega_\Lambda = 0.69$, and $H_0 = 67.7\,\mathrm{km\,s^{-1}Mpc^{-1}}$, as presented by the Planck Collaboration et al. (2020).

Most of works on Radio luminosity functions (RLFs) have used luminosity values at 1.4 GHz (e.g. Mauch and Sadler, 2007; Simpson et al., 2012; McAlpine et al., 2013; Šlaus et al., 2020). In order to have the opportunity to compare our results with previous literature, we will convert our EMU luminosities to be at that frequency. Using Eq. 1.5, we can obtain luminosities at 1.4 GHz from our values at 944 MHz. The distribution of such luminosities, as a function of predicted photometric redshift, is presented for both samples, in Fig. 9.5.

Attending to the intrinsic uncertainties of the original pipeline and the changes introduced to the models in the extended prediction pipeline, the luminosity distributions of Fig. 9.5 appear to be relatively clean. Two potential issues can be, nevertheless, identified. First, a fraction of luminosities are located below the $5-\sigma$ detection limit of EMU-PS. The presence of these populations (both galaxies and AGN) can be explained by two interrelated factors. One of them

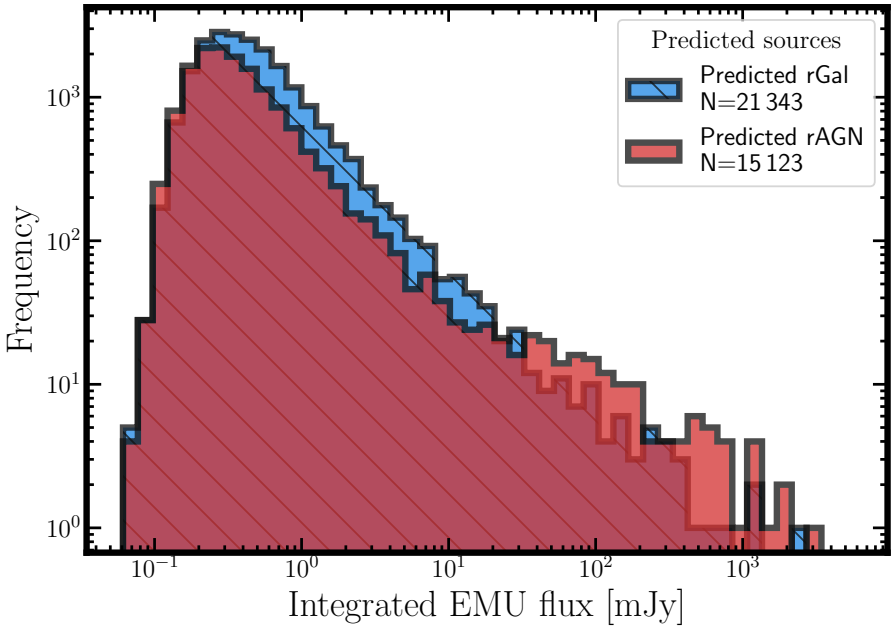


Figure 9.4: Distribution of EMU integrated fluxes (at 944 MHz) for, in hatched blue, predicted radio-detectable galaxies and, in red, predicted radio-detectable AGN in the area of the EMU-PS catalogue.

is that the detection limit presented in Fig. 9.5 has been calculated with the mean detection depth of the survey (25 µJy/beam) and the other is that the detection limit used the fixed spectral index $\alpha = -0.7$, while the detected sources have been analysed with their own spectral indices.

Another potential issue that can be identified is related, in the case of predicted radio galaxies, to the sources with luminosity values that are too high to be univocally defined as being originated in star-forming processes. At low luminosity ranges, such as those reached by the latest radio observatories and surveys, it becomes difficult to differenciate between radio emission from AGN and that from SF. One simple approach to separate both populations, and to classify sources with faint radio luminosities, is using a single value for which all sources brighther that that might be labelled as AGN. For instance, one value used is 10^{25} W Hz^{-1}, above which sources can be considered, without large uncertainties, as radio-loud AGN (e.g. Williams and Röttgering, 2015; Mo et al., 2020).

Expanding on the idea of setting a threshold, it is possible to obtain a function for this limit that might depend on, for instance, redshift values. Given that the distributions of AGN and galaxy luminosities (i.e. LFs) have different behaviours, it is expected that the curves of both values will cross at some point (Magliocchetti, 2022). Following the results from Magliocchetti et al. (2014), which have been based upon the work of Magliocchetti et al. (2002), Mauch and Sadler (2007), and McAlpine et al. (2013), a threshold can be defined as a function of the redshift values of the sources. The expression can be written as follows:

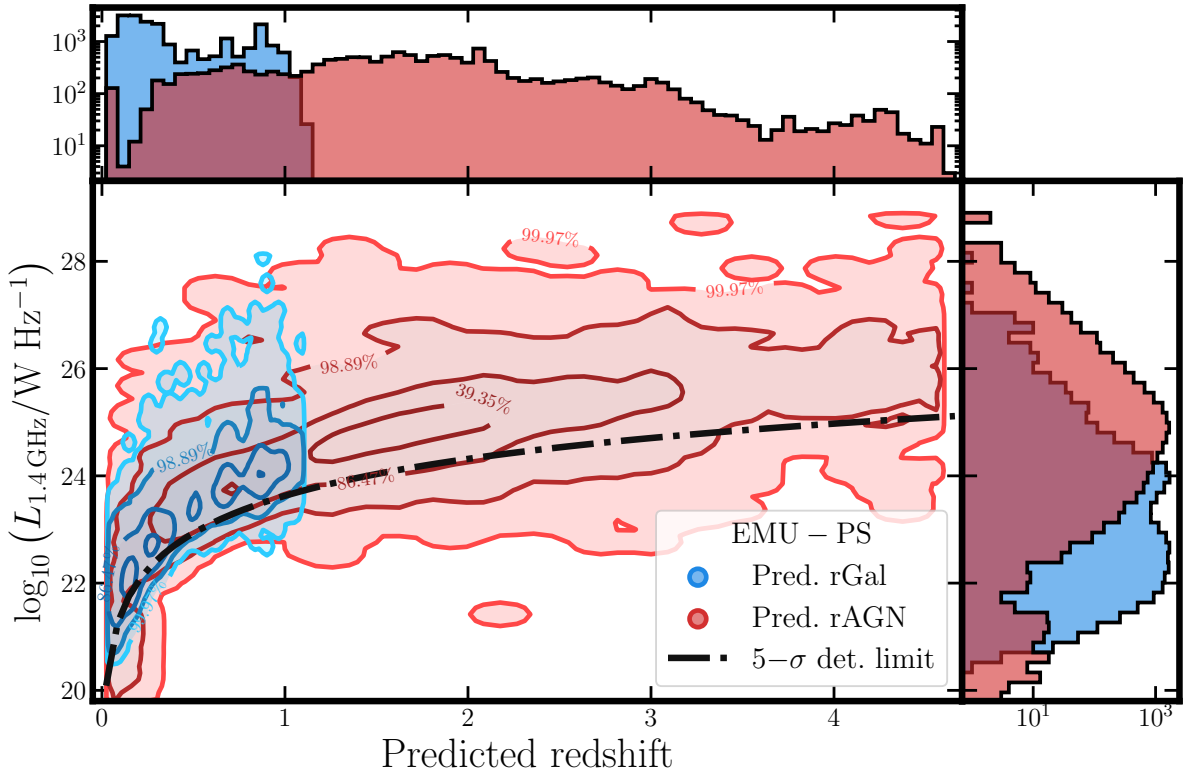


Figure 9.5: Distribution of predicted 1.4 GHz luminosities vs predicted photometric redshifts for radio-detectable AGN (in red) and radio-detectable galaxies (blue) in the area of the EMU-PS catalogue. In the top panel, the histograms of the predicted redshifts for both populations are presented. In the right-hand side of the figure, the distributions of predicted 1.4 GHz luminosities are displayed following the color code of the legend. In the central panel, two contour plots represent the joint distribution of 1.4 GHz luminosities and predicted photometric redshifts of both predicted radio-detectable AGN and radio-detectable galaxies. Contours represent the 1, 2, 3 and 4 – σ two-dimensional levels of the distribution (39.35 %, 86.47 %, 98.89 % and 99.97 %, respectively, of the corresponding sample). Black, dashed line represents the luminosity of a source at five times the noise level of the EMU-PS survey (25 m Jy). That is to say, it represents the detection limit of the EMU-PS catalogue.

$$\log_{10}\left(L_{\rm cross}\right) = \log_{10}\left(L_{0,\rm cross}\right) + z\,, \tag{9.4}$$

in which $L_{0,\rm cross} = 5.01 \times 10^{21}\,\mathrm{W\,Hz^{-1}\,sr^{-1}}$ for $z \leq 1.8$, and $L_{0,\rm cross} = 3.16 \times 10^{23}\,\mathrm{W\,Hz^{-1}\,sr^{-1}}$ for $z > 1.8$. After the addition of the angular factor 4π, these values can be expressed as $L_{0,\rm cross} = 6.3 \times 10^{22}\,\mathrm{W\,Hz^{-1}}$ for $z \leq 1.8$ and $L_{0,\rm cross} = 3.97 \times 10^{24}\,\mathrm{W\,Hz^{-1}}$ for $z > 1.8$. We decided to use these thresholds in our EMU-PS sample to alleviate the existence of too-bright radio galaxies. Thus, all predicted radio galaxies that presented a 1.4 GHz luminosity above the mentioned limit were re-labelled as predicted AGN and the corresponding branch of the prediction pipeline is applied to them. That is, a new photometric redshift value is predicted for all of them. The modified distributions of luminosities and redshifts of the re-labelled AGN and galaxies sets are presented in Fig. 9.6.

It is possible to see, in Fig. 9.6, that the locus, in the (z, L) plane, of predicted radio-detected galaxies is much smaller than previously (see Fig. 9.5). The new distribution is consistent with

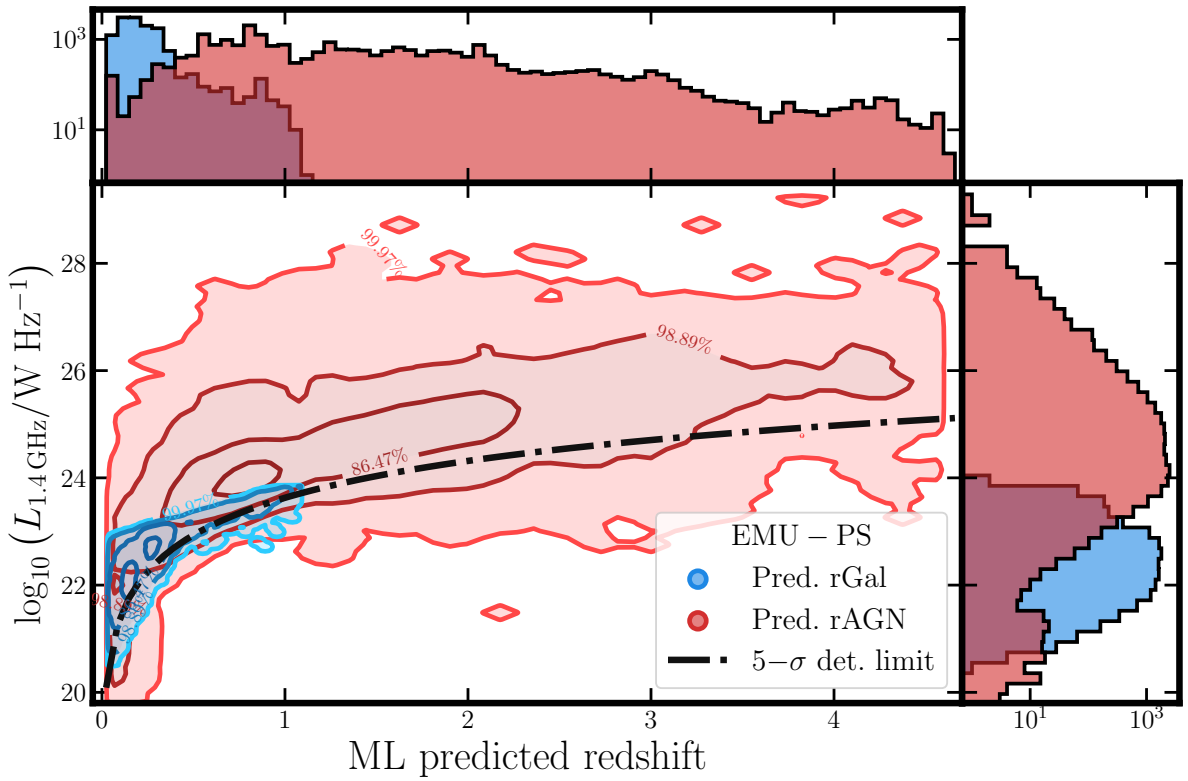


Figure 9.6: Distribution of predicted 1.4 GHz luminosities vs predicted photometric redshifts for radio-detectable AGN (in red) and radio-detectable galaxies (blue) in the area of the EMU-PS catalogue. Galaxies with high 1.4 GHz luminosities have been re-labelled as AGN, following the prescriptions by Magliocchetti et al. (2014) and Magliocchetti (2022). Description as in Fig. 9.5.

the expectation that, in general, radio emission from AGN is stronger than that of SF episodes. After the re-labelling, 12 561 sources are classified as predicted radio galaxies, and 23 712, as predicted AGN.

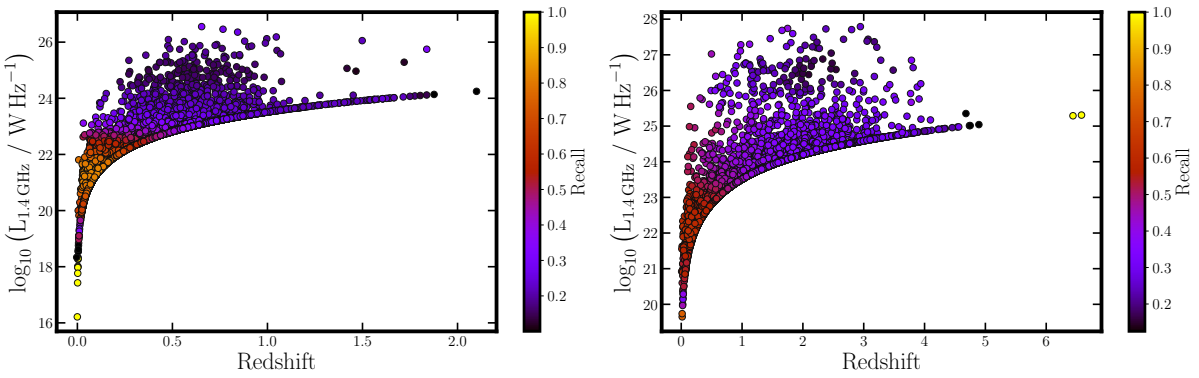
In order to obtain RLFs for our predicted radio-detected AGN, and from the description in Sect. 1.1, the Python package `KDELF`[3] (Yuan et al., 2022; Yuan et al., 2020) was used. Among other techniques, it can derive LFs using an adaptive Kernel density estimation (KDE) approach (Abramson, 1982; Davies et al., 2018b) as well as using a binning estimator, as presented by Page and Carrera (2000).

The KDE-based approach allows to optimise the bandwidths of the kernels. The functional form of the LF is:

$$\hat{\phi}_a(z, \mathrm{L}) = \frac{N_{\mathrm{eff}}(Z_2 - Z_1)\hat{f}_{wa}(x, y|h_1, h_2, \beta)}{(z - Z_1)(Z_2 - z)\Omega\dfrac{\mathrm{dV}}{\mathrm{d}z}}, \tag{9.5}$$

where Z_1 and Z_2 are the redshift limits of the studied sample, $\hat{f}_{wa}(x, y|h_1, h_2, \beta)$ corresponds to

[3] https://github.com/yuanzunli/kdeLF



(a) True radio galaxies

(b) True radio AGN

Figure 9.7: True 1.4 GHz luminosity vs redshift for (a) radio galaxies and (b) radio AGN in the HETDEX catalogue.Sources are coloured according to their estimated recall values and following each individual colourbar.

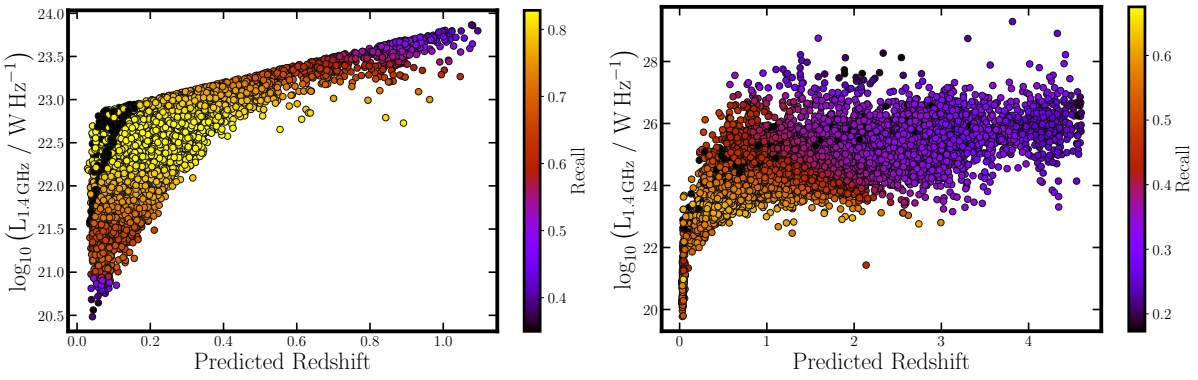
the density function of the pair (x, y) which is the equivalent of (z, L) in the KDE space. $N_{\mathrm{eff}} = \sum_{i=1}^{n} w_i$ is the effective size of such sample. The weight value is defined as $w_i \equiv 1/\mathcal{P}(z_i, L_i)$ with $\mathcal{P}$ the selection function.

The selection function, $\mathcal{P}(z, L)$, summarises the corrections that the distribution of sources must suffer in order to be as close as possible to our best guess of their real distribution in the space of redshifts and luminosities. Taking advantage of the use of ML predictions, it is possible to obtain a correction for completeness (recall) in our selected sample that can be added to the selection function. In order to obtain these values, we took all known radio-detected sources (i.e. AGN and galaxies) in the HETDEX field as well as their predicted class, redshift values, and estimated 1.4 GHz luminosities (assuming a spectral index $\alpha = -0.7$). For each element in the $\left(z_{\mathrm{pred}}, \log_{10}\left(L_{1.4\,\mathrm{GHz}}\right)\right)$ plane, all the sources located within a dimensionless distance of 1 are selected (i.e. their nearest neighbours). For these subsets of known radio sources, the class recall (cf. Eq. 3.4) is calculated. Their distribution of values is presented in Fig. 9.7.

Then, each of the predicted sources from the EMU-PS catalogue is placed in the same plane of their corresponding class among the HETDEX sources. The ten closest HETDEX sources, with a Euclidean distance, are selected and their recall values are averaged with the inverse of their distance to the EMU-PS source as weights. This averaged value is assigned to the predicted source as their estimated recall. The distributions of such values are presented in Fig. 9.8.

Having obtained the completeness values for the predicted sources among the EMU-PS sources, a first definition for the selection function is the following:

$$\mathcal{P}(z_i, L_i) = \mathrm{Completeness}_{\mathrm{ML}}\left(z_i, L_i\right) , \tag{9.6}$$



(a) Predicted radio galaxies

(b) Predicted radio AGN

Figure 9.8: Predicted 1.4 GHz luminosity vs predicted redshift for predicted (a) radio galaxies and (b) radio AGN in the EMU-PS catalogue.Sources are coloured according to their estimated recall values and following each individual colourbar.

with the ML-based completeness ranging between 0 and 1. It can be seen, then, as the completeness decreases, the selection function follows the same behaviour. Additional factors might be included in the definition of the selection function. Some of them involve a correction for the resolution bias (e.g. Prandoni et al., 2001; Prandoni et al., 2018; Mandal et al., 2021) and a correction for the Eddington bias (Eddington, 1913; Eddington, 1940). For our project, and since it is more a proof of concept rather than a full analysis, only a completeness correction will be applied to the studied sources.

Taking into account the distribution of predicted redshifts from our set of predicted radio-detected AGN and radio-detected galaxies and following the work by van der Vlugt et al. (2022), 11 redshift bins were created: $(0.01, 0.1]$, $(0.1, 0.4]$, $(0.4, 0.6]$, $(0.6, 0.8]$, $(0.8, 1.0]$, $(1.0, 1.3]$, $(1.3, 1.6]$, $(1.6, 2.0]$, $(2.0, 2.5]$, $(2.5, 3.3]$, and $(3.3, 4.8]$. In the case of radio-detected galaxies, the first six redshift bins contain sources of this type (i.e. up to $z = 1.20$). The number of sources in each redshift bin is shown in Table 9.3.

We will obtain LF values from the use of the adaptive parametric method as well as through binned estimation in each predicted redshift bin. For the calculation of the binned LF, a bin of size $\Delta \log_{10} L = 0.3$ was chosen following, for instance, Ross et al. (2013), Kondapally et al. (2022), Yuan et al. (2022), and Alqasim and Page (2023). For the KDE estimation of the LF, the package `KDELF` implements a fully Bayesian Markov Chain Monte Carlo (MCMC) determination of the posterior distributions of the parameters of the LF (h_1, h_2, and β). From this implementation, we can obtain median values as well as their uncertainties (in a Bayesian context, credible intervals). Binned and adaptive LFs, for each redshift bin, are displayed in Fig. 9.9. Additionally, Figs. 9.10 and 9.11 shows an integrated view of all values of the adaptive KDE LF in several redshift bins simultaneously.

Table 9.3: Number of predicted radio-detectable sources (AGN and galaxies) in the EMU-PS area by predicted resdhift bin.

z bin	Radio AGN	Radio galaxies
$0.01 < z \leq 0.1$	162	2097
$0.1 < z \leq 0.4$	721	9589
$0.4 < z \leq 0.6$	2281	418
$0.6 < z \leq 0.8$	2911	197
$0.8 < z \leq 1.0$	4030	235
$1.0 < z \leq 1.3$	2458	25
$1.3 < z \leq 1.6$	2630	...
$1.6 < z \leq 2.0$	3430	...
$2.0 < z \leq 2.5$	2513	...
$2.5 < z \leq 3.3$	1938	...
$3.3 < z \leq 4.8$	638	...

In the case of the binned LF, it can be seen that, for higher luminosities, it presents a noticeable raise. This effect has been discussed in previous works. For instance, Miyaji et al. (2001), Croom et al. (2009), and Palanque-Delabrouille et al. (2016) have noted that the method by Page and Carrera (2000, binned LF), might cause pronounced biases. The choice of the specific point in the $\log(L)$ bin can seriously impact in the final estimated value. Additionally, by assuming a uniform distribution in every luminosity bin, this method is unable to correct for the incompleteness in each individual sub-range. This last issue can be of the utmost relevance in the brightest end of the LF, where the number of sources decreases significantly.

9.3 Radio counterpart assessment

One of the main issues of multi-wavelength studies of astrophysical sources is the identification of counterparts in observations from different instruments and bands (cf. Sect. 2.3). Accurate positioning of sources is crucial to allow further studies (e.g. spectroscopic targeting). While optical and IR observations have reached sub-arcsecond positional accuracy (e.g. *WISE*, Wright et al., 2010; PS1, Chambers et al., 2016), deep radio surveys have only recently obtained such resolutions (e.g. Sweijen et al., 2022; Ye et al., 2023, for the LOFAR radio telescope). Thus, most of radio identifications lack accurate positional measurements needed for precise counterpart identification and to reduce uncertainties in the use of SED modelling.

There are a few approaches to find and link multi-wavelength, multi-instrument counterparts of sources. One of the simplest way to find such sources is through direct cross match between catalogues. A search radius is defined and centred in the position of the source for

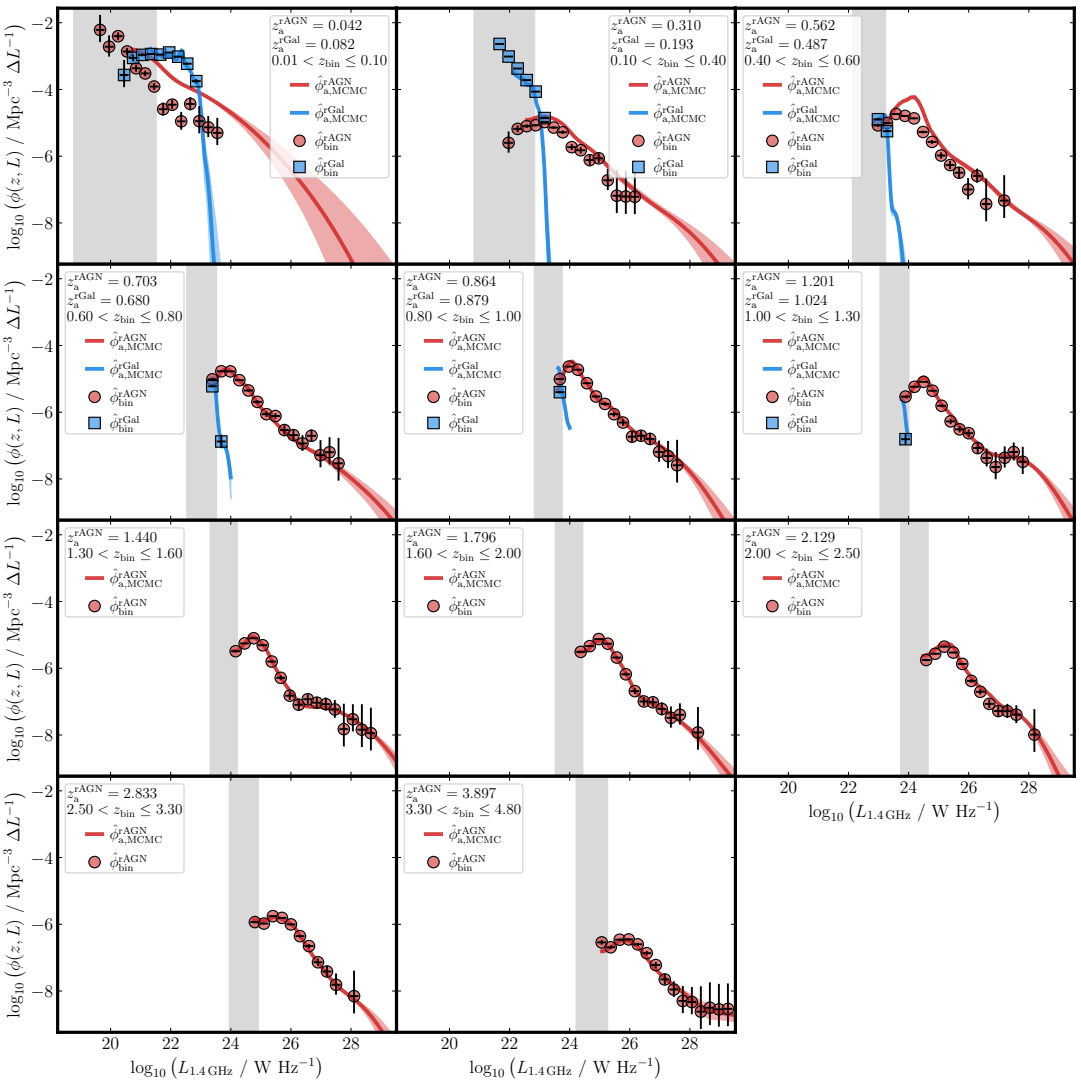


Figure 9.9: RLF (1.4 GHz) in EMU-PS binned by predicted z values. In red, values for predicted radio AGN and, in blue, for predicted radio galaxies. Circles and squares show binned RLF with $1-\sigma$ uncertainty error bars. Solid lines show the MCMC-sampled adaptive KDE RLF with $3-\sigma$ credible interval shaded regions. Grey regions show 1 to $5-\sigma$ detection levels from EMU-PS measurements.

which counterparts are needed. Then, all sources in the target catalogue located inside the circle of the previously defined radius are considered candidate counterparts. Depending on the conditions of the problem and the used catalogues, one of these candidates can be selected as the proper counterpart (e.g. the closest source either in angular or geometrical distance). This thesis is an example of the use of this technique (cf. Sect. 5.3). Other instances of the use of direct cross match of catalogues are Barbieri and Bertola (1972), Agüeros et al. (2005), Bianchi et al. (2007), Drake et al. (2014), Norris et al. (2021), and Storey-Fisher et al. (2023).

When postional errors are large or PSF or synthesised beams are large enough to have several sources from the other catalogue inside it, direct cross matching cannot be used. A more advanced approach is that of Maximum Likelihood Radio (MLR; Richter, 1975; de Ruiter et al.,

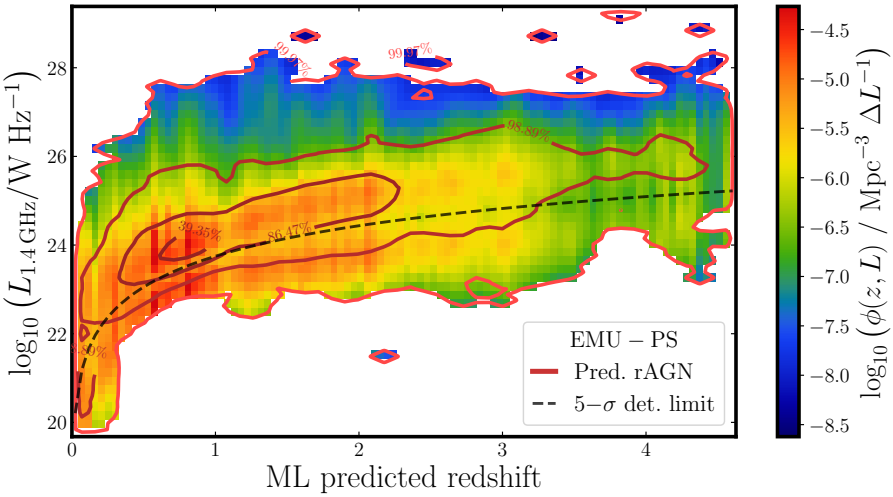


Figure 9.10: Unified RLF (1.4 GHz) for predicted radio AGN in EMU-PS. Contours represent the 1, 2, 3 and 4 − σ two-dimensional levels of the distribution of sources. For each bin in the (1.4 GHz, z) plane, the color represents the value of the adaptive KDE RLF following the colour coding in the colourbar. Black, dashed line depicts the 5 − σ detection limit from EMU-PS.

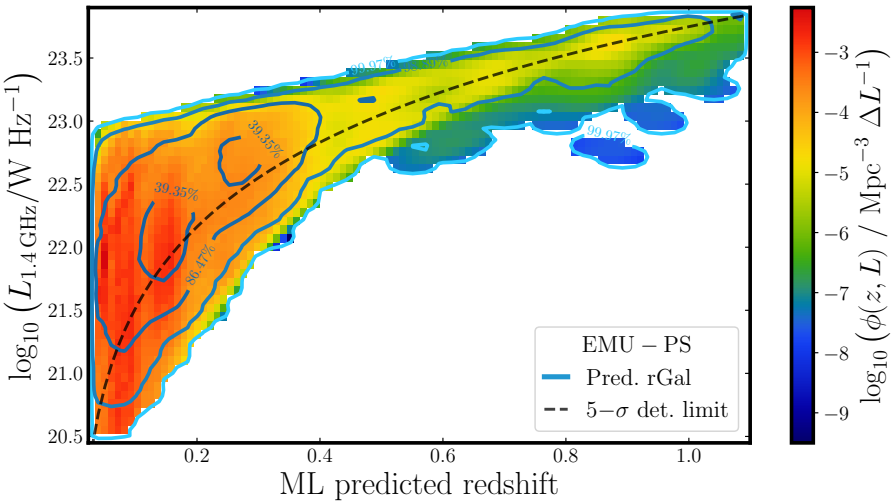


Figure 9.11: Unified RLF (1.4 GHz) for predicted radio galaxies in EMU-PS. Description as in Fig. 9.10.

1977; Prestage and Peacock, 1983; Wolstencroft et al., 1986; Sutherland and Saunders, 1992). As defined by Sutherland and Saunders (1992), it looks for the sources that optimise the ratio of the likelihoods of being a genuine counterpart over that of being a background candidate. Theses likelihoods depend on the density of sources in both catalogues, their magnitude distributions, and their positional errors (Brusa et al., 2007). One advantage of this technique is that it can output the degree of reliability of each counterpart allowing the researcher to select, if needed, the most secure sources. Some examples of the application of the MLR method include Brusa et al. (XMM-COSMOS; 2007), Abdo et al. (Fermi; 2010), Xue et al. (Chandra Deep Field-South; 2011), LaMassa et al. (Stripe 82X; 2016), Marchesi et al. (Chandra COSMOS; 2016), Ananna et al. (Stripe 82X; 2017), Auge et al. (2023), Hardcastle et al. (LoTSS; 2023), and Whittam et al.

(MIGHTEE-COSMOS; 2024).

A third method is the Bayesian approach. Contrary to the previous techniques, it does not rely on the specific distribution of sources in the studied catalogues, using Bayesian priors to derive the most likely counterparts of the base catalogue. In this way, it does not suffer from being applied to small areas (Salvato et al., 2018). This method was first introduced by Budavári and Szalay (2008) and it can be applied to the search for counterparts in simultaneous catalogues.

Additionally, ML can be used to derive the most likely counterpart of sources in catalogues. By using photometric information (or other properties) from sources detected in other wavelengths, it is possible to train a model and extract the probability of that source to have a counterpart in a new catalogue. One early example of such technique is the work by Rohde et al. (2005) and Rohde et al. (2006) where the authors used SVMs, together with model calibration (see Sect. 3.4) in order to obtain a counterpart probability. More recently, Liu et al. (2019) used GP modelling to quantify the confidence of associations of Atacama large millimeter/submillimeter array (ALMA) detections in the COSMOS fields. Furthermore, Schneider et al. (2022) used SVMs to extract stellar counterparts of sources in the eROSITA Final Equatorial Depth Survey (eFEDS; Brunner et al., 2022).

In particular, the radio astronomical community has forecast that the use of ML will be instrumental to find radio counterparts of large catalogues (Lazio et al., 2014). Nevertheless, and as stated in Sect. 8.1.2, the number of works using ML for predicting the radio detectability of sources (problem comparable to finding radio counterparts) is, to date, very low.

The lack of such studies prompts us to investigate the potential use of ML techniques to quantify the probability of a detection to have a counterpart in a different photometric catalogue. In order to assess the idea, we will utilise the radio-detection classification model described in Sect. 9.2 and Appendix B (which is a modification of the model presented in Chapters 4 and 6). Following the discussion of Rohde et al. (2006), we can make directly use of the output probabilities given by the models since they have been calibrated and their distribution is well behaved.

The pipeline described in Sect. 9.2 and Appendix B was applied to the IR-detected sources in the EMU-PS area. In order to test the radio counterparts, we selected the sources, regardless of their initial classification, that were predicted to be radio-detectable AGN. In particular, we focused on the sources that presented high probabilities to be of such class (i.e. probability of being AGN higher than $P(\mathbb{C}) = 0.7$ and probability to have radio detection higher than

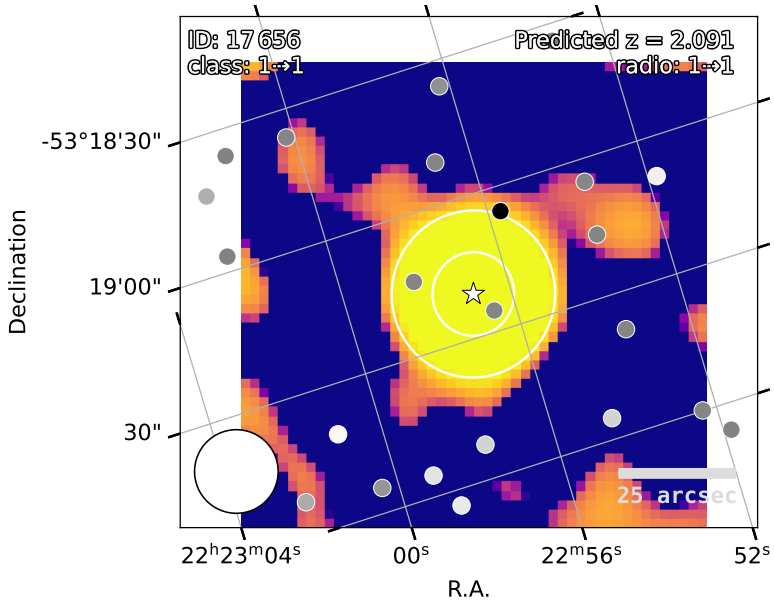
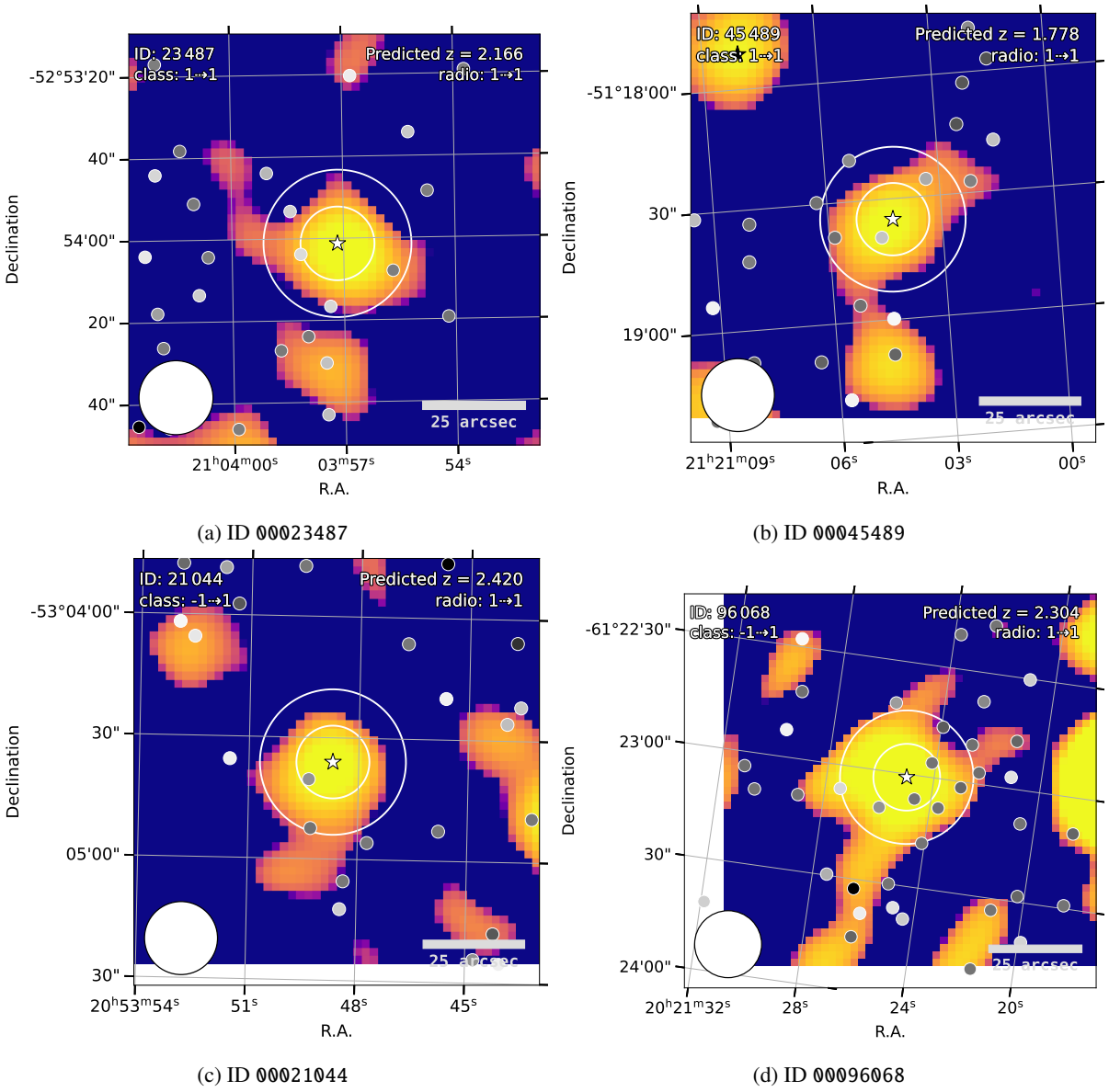


Figure 9.12: Postage stamp of CW-detected source ID `00017656` in the EMU-PS area (located in the centre of the image). All small circles show the postion of a CW detection in the displayed region without a radio counterpart, while all stars are CW detections with a confirmed radio detection. Face colours of stars and small circles correlate with the predicted probability of such CW detections to have a radio detection associated to them, where a brighter hue represents a higher probability. For displaying purposes, all emission below a $1-\sigma$ detection limit has been set to zero (0). Concentric white circles in the centre of the image limit two regions with a radius of 1 and 2 Full width at half maximum (FWHM) centred in the selected source. White circle in bottom right corner represents the EMU-PS synthesised beam size and the gray horizontal lines in the bottom right corner denote a $25''$ scale. In the top left corner of figure, the identification number of source is written, as well as its confirmed and predicted AGN (i.e. `class`) states, where `0` represents a galaxy, 1 represents an AGN, and -1, a source without prior identification. In the top right corner of the figure, the predicted redshift of the source is included, as well as its confirmed and predicted radio detection (i.e. `radio_detect`) status, with `0` representing a source without radio detection (confirmed or predicted) and 1 stands for radio detection.

$P(\mathbb{C}) = 0.8$). Then, we plotted these predicted radio-detectable sources on top of the map of EMU-PS together with all other IR-detections in the surrounding region. Some sources from this selection are shown in Figs. 9.12 to 9.15a.

It is worth noting that in each image, all IR-detected sources have been plotted, regardless of their initial or predicted class. The inclusion of the full sample implies that the radio prediction model was applied to all candidates and thus, more uncertainties have been included in the results from its application. This uncertainties do not hinder the analysis regarding counterparts.

A first example is shown in Fig. 9.12. It presents a radio-detected AGN (ID `00017656`) that has been predicted to have be in the same category (i.e. radio-detectable AGN, an accurate prediction). The star in the middle of the field shows that the source has been detected in the EMU-PS data, as shown by the bright region in the background image. There are two additional CW-detected sources within one synthesised beam of distance from the selected target. They present a darker face colour, indicating that their probability of having a radio counterpart is lower than that of `00017656`. Thus, the use of the pipeline with the selected source shows that it can be inferred that the large radio source in the background (and only associated by means of



(a) ID `00023487`

(b) ID `00045489`

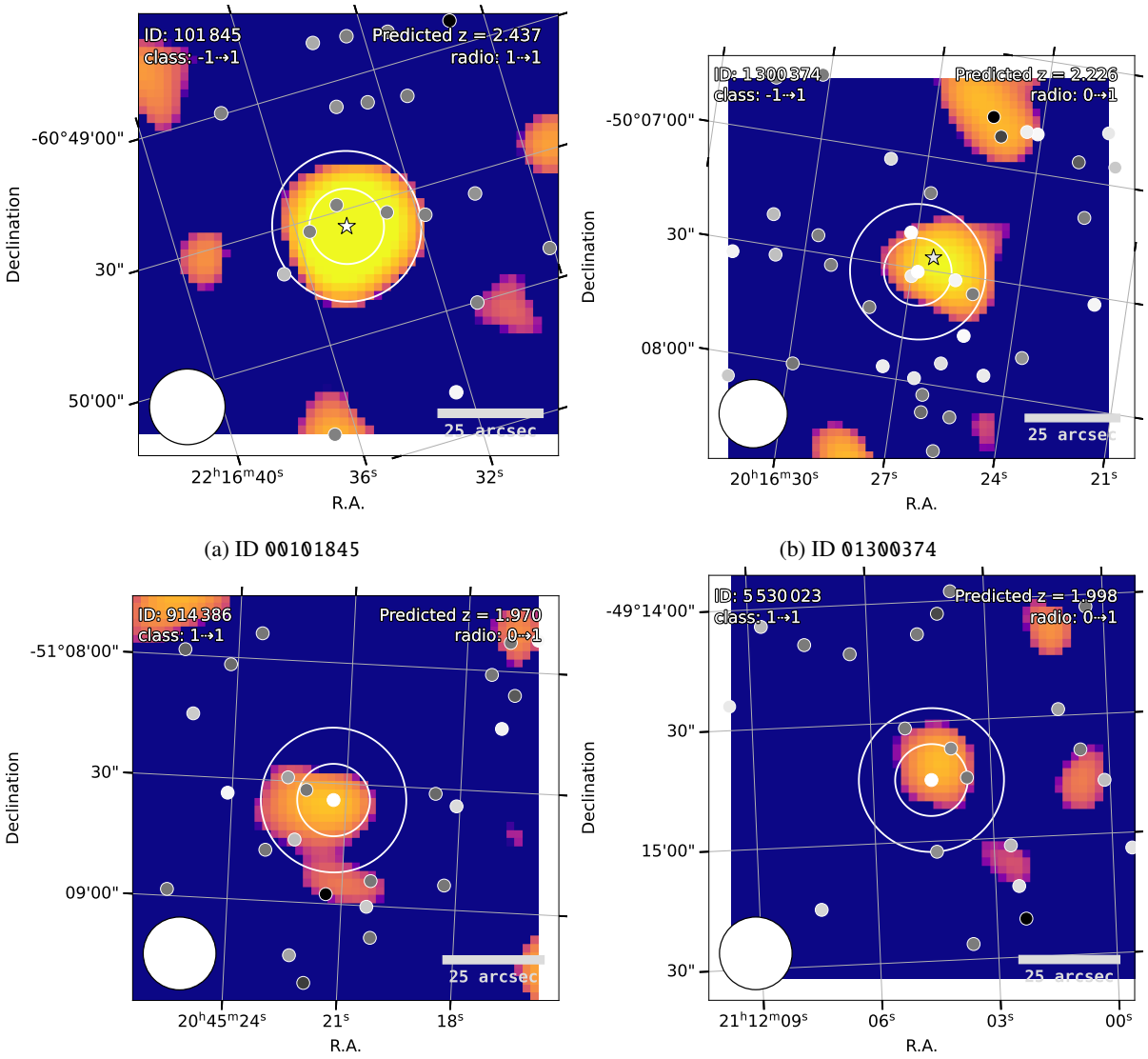
(c) ID `00021044`

(d) ID `00096068`

Figure 9.13: Postage stamp of CW-detected sources IDs `00023487`, `00045489`, `00021044`, and `00096068` in the EMU-PS area. Description and details as in Fig. 9.12.

distance) can be the counterpart of `00017656` with a high likelihood. What Fig. 9.12 presents, then, is the expected output of our prediction pipeline.

The same behaviour can be seen in Figs. 9.13a, 9.13b, 9.13c, 9.13d, and 9.14a. In particular, Figs. 9.13a and 9.13b show two confirmed radio-detected AGN that have been correctly predicted surrounded by several CW sources with lower probability of having radio counterparts. Interestingly, the radio source in Fig. 9.13b resembles the emission of a central source with two lobular arms, as radio AGN with a bent jet. Then, Figs. 9.13c, 9.13d, and 9.14a depict EMU-PS confirmed detections that do not have an associated AGN or galaxy confirmation. Nevertheless, they have been predicted as AGN with high confidence. As previously, these predictions have the highest likelihood of being detected in the radio in their



(a) ID `00101845`

(b) ID `01300374`

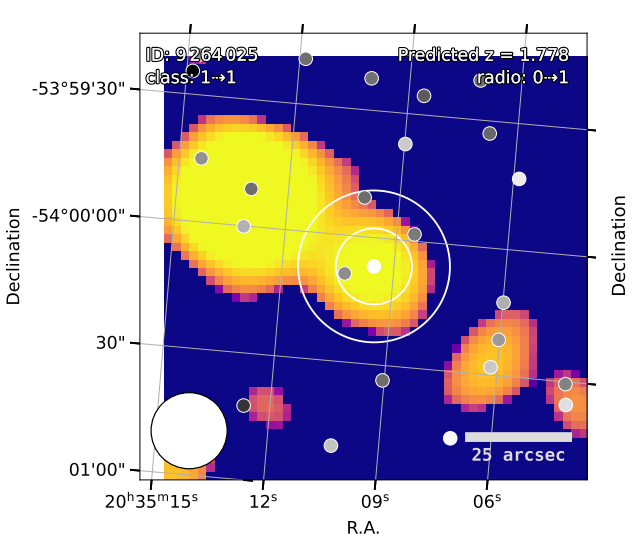
(c) ID `00914386`

(d) ID `05530023`

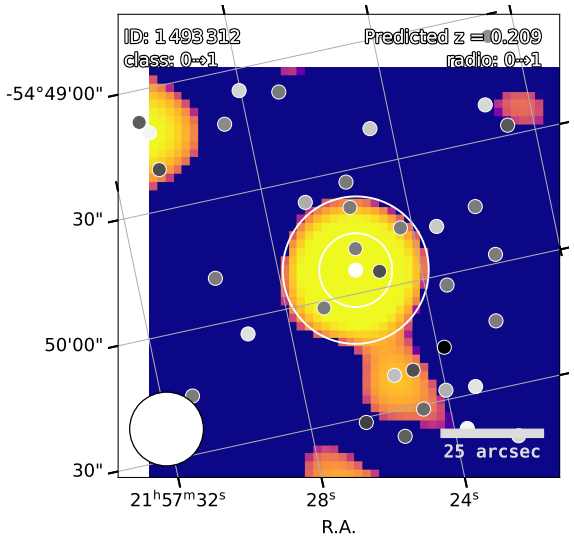
Figure 9.14: Postage stamp of CW-detected sources IDs `00101845`, `01300374`, `00914386`, and `05530023` in the EMU-PS area. Description and details as in Fig. 9.12.

surroundings (within two FWHM from their position). It is important to note that the emission depicted in Fig. 9.13d might not remind the reader of an AGN or a similar source. Even though the radio prediction and the cross-match are in agreement, further studies are needed to confirm the nature of such source.

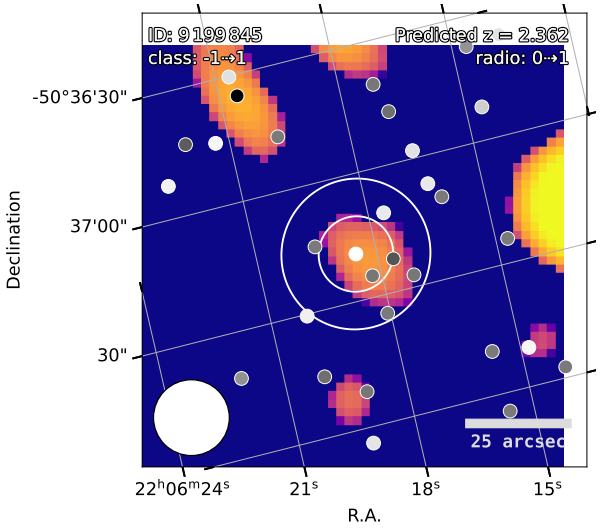
A different scenario is shown in Figs. 9.14b, 9.14c, 9.14d, 9.15a, 9.15b, and 9.15c, where all CW-detected sources could not be associated (with a direct cross-match) to a source in EMU-PS. It is important to note that all emission shown in the postage stamps is above a $1-\sigma$ limit, while the detections listed in the EMU-PS catalogue are above the $5-\sigma$ limit (Norris et al., 2021). Thus, it is possible that the catalogue misses faint sources that are picked and predicted by our pipeline.



(a) ID 09264025



(b) ID 01493312



(c) ID 09199845

Figure 9.15: Postage stamp of CW-detected sources IDs 09264025, 01493312, and 09199845 in the EMU-PS area. Description and details as in Fig. 9.12.

Figure 9.14b presents an example worth noting. Source ID 01300374, which does not present any prior classification, shows the highest probability of radio detection in its surroundings, but there is an additional source that has been labelled as the counterpart to the background radio source. This source, shown as a star with a darker face colour, has obtained a lower probability to be detected in the radio bands. Therefore, our detection pipeline has contested the radio counterpart assignation by distance. Additional measurements are needed to determine the most likely counterpart.

Sources ID 00914386 and ID 05530023, depicted in Figs. 9.14c and 9.14d, have been originally labelled as an AGN without any radio detection. Our pipeline has assigned to them a high probability to have a radio counterpart over their neighbouring sources. At the same time,

their background images show a bright source in the same position. Thus, we have predicted that these source can have a radio counterpart (which might have been previously catalogued should EMU-PS have used a lower detection threshold). A similar situation can be seen in Fig. 9.15a with source ID `09264025`. Its main difference is that it is located in what might be a more radio-populated region that could present some overlapping sources.

Finally, two sources without an AGN label are presented. Source ID `01493312` (Fig. 9.15b), without a prior label as galaxy, is located in a very IR-crowded region and it has been given, by our prediction pipeline, the highest radio-detection likelihood over their neighbours. In the background EMU-PS image, it is possible to see bright radio emission from a point-like source that has a bright appendix. Thus, we have been able to predict the existence of radio emission that was missed by the EMU-PS catalogue. Our last example source is ID `09199845` in Fig. 9.15c. Without any prior label (`class` or `radio_detect`), it has been predicted to be a radio-detectable AGN among all its neighbouring detections. The EMU-PS image shows a relatively bright source that might be associated to our prediction.

All these examples show that our prediction pipeline can be used, without any further modification, to understand the location and distribution of radio counterparts of IR-detected sources. By virtue of its training with very deep radio observations (i.e. LoTSS), our pipeline can also help finding sources that might have been overlooked or discarded by other radio surveys and catalogues, as was the case with the EMU-PS catalogue.

This page intentionally left blank.

10

Future developments

The results presented here allow us to look for further developments. We list here some steps that can be taken in order to expand our knowledge on either the prediction pipeline or the results obtained from its use.

10.1 Extensive feature importance analysis

Expanding on the analysis of Chapter 8, the behaviour of the model, when applied to different data sets, can be studied extensively.

10.2 Evolutionary Map of the Universe

Starting with Sects. 9.2 and 9.3, which show possible applications of the prediction pipeline on data from the EMU-PS, more studied can be performed on the areas covered by EMU. The full products from the EMU survey can be subject to our prediction pipeline and as a way to obtain a large number of radio-AGN candidates. The main barrier to achieve such goal is related to the existence of coverage from deep and homogeneous optical surveys. Our training data incorporates measurements from PS1, which is not fully available in the southern hemisphere. For the EMU-PS data, measurements from the

10.3 Square Kilometre Array

Ultimately, our prediction pipeline can be applied to the full area covered by the future SKA in order to accelerate the detection of AGN.

This page intentionally left blank.

Summary

This is the summary of the text we have produced.

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.

- ML models cannot be only used for a direct prediction of a value (or a set of values). 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. the Rapid ASKAP Continuum Survey -RACS-, EMU, and theMeerKAT International GHz Tiered Extragalactic Exploration -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 and 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 infrared (FIR), X-ray, and multi-survey radio measurements makes part of our efforts to improve detections, not only in radio, but in additional wavelengths.

This page intentionally left blank.

Data and software acknowledgements

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[1]) 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 Ana-

[1] https://lofar-surveys.org/

lysis 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 scientific work uses data obtained from Inyarrimanha Ilgari Bundara / the Murchison Radio-astronomy Observatory. We acknowledge the Wajarri Yamaji People as the Traditional Owners and native title holders of the Observatory site. CSIRO's ASKAP radio telescope is part of the Australia Telescope National Facility (https://ror.org/05qajvd42). Operation of ASKAP is funded by the Australian Government with support from the National Collaborative Research Infrastructure Strategy. ASKAP uses the resources of the Pawsey Supercomputing Research Centre. Establishment of ASKAP, Inyarrimanha Ilgari Bundara, the CSIRO Murchison Radio-astronomy Observatory and the Pawsey Supercomputing Research Centre are initiatives of the Australian Government, with support from the Government of Western Australia and the Science and Industry Endowment Fund.

Part of this work is based on data obtained from the ESO Science Archive Facility with

DOI(s): https://doi.org/10.18727/archive/56.

This project used public archival data from the Dark Energy Survey (DES). Funding for the DES Projects has been provided by the U.S. Department of Energy, the U.S. National Science Foundation, the Ministry of Science and Education of Spain, the Science and Technology Facilities Council of the United Kingdom, the Higher Education Funding Council for England, the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign, the Kavli Institute of Cosmological Physics at the University of Chicago, the Center for Cosmology and Astro-Particle Physics at the Ohio State University, the Mitchell Institute for Fundamental Physics and Astronomy at Texas A&M University, Financiadora de Estudos e Projetos, Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro, Conselho Nacional de Desenvolvimento Científico e Tecnológico and the Ministério da Ciência, Tecnologia e Inovação, the Deutsche Forschungsgemeinschaft, and the Collaborating Institutions in the Dark Energy Survey. The Collaborating Institutions are Argonne National Laboratory, the University of California at Santa Cruz, the University of Cambridge, Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas-Madrid, the University of Chicago, University College London, the DES-Brazil Consortium, the University of Edinburgh, the Eidgenössische Technische Hochschule (ETH) Zürich, Fermi National Accelerator Laboratory, the University of Illinois at Urbana-Champaign, the Institut de Ciències de l'Espai (IEEC/CSIC), the Institut de Física d'Altes Energies, Lawrence Berkeley National Laboratory, the Ludwig-Maximilians Universität München and the associated Excellence Cluster Universe, the University of Michigan, the National Optical Astronomy Observatory, the University of Nottingham, The Ohio State University, the OzDES Membership Consortium, the University of Pennsylvania, the University of Portsmouth, SLAC National Accelerator Laboratory, Stanford University, the University of Sussex, and Texas A&M University. Based in part on observations at Cerro Tololo Inter-American Observatory, National Optical Astronomy Observatory, which is operated by the Association of Universities for Research in Astronomy (AURA) under a cooperative agreement with the National Science Foundation.

This work has made use of data from the European Space Agency (ESA) mission *Gaia* (https://www.cosmos.esa.int/gaia), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement.

The Legacy Surveys imaging of the DESI footprint is supported by the Director, Office

of Science, Office of High Energy Physics of the U.S. Department of Energy under Contract No. DE-AC02-05CH1123, by the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility under the same contract; and by the U.S. National Science Foundation, Division of Astronomical Sciences under Contract No. AST-0950945 to NOAO.

This research has made use of NASA's Astrophysics Data System, TOPCAT[2] (Taylor, 2005), JupyterLab[3] (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`[4] (`v2.3.10`; Ali, 2020), `scikit-learn` (`v0.23.2`; Pedregosa et al., 2011), `pandas`[5] (`v1.4.2`; McKinney, 2010), `Astropy`[6], a community-developed core Python package for Astronomy (`v5.0`; Astropy Collaboration et al., 2013; Astropy Collaboration et al., 2018; Astropy Collaboration et al., 2022), `Matplotlib` (`v3.5.1`; Hunter, 2007), `betacal`[7] (`v1.1.0`), `CMasher`[8] (`v1.6.3`; van der Velden, 2020), and `faiss`[9] (`v1.7.2`; Johnson et al., 2019).

[2] http://www.star.bris.ac.uk/~mbt/topcat/
[3] https://jupyter.org
[4] https://pycaret.org
[5] https://pandas.pydata.org
[6] https://www.astropy.org
[7] https://betacal.github.io
[8] https://github.com/1313e/CMasher
[9] https://faiss.ai

References

Abbott, T. M. C., Abdalla, F. B., Allam, S., et al. (Dec. 2018). 'The Dark Energy Survey: Data Release 1'. In: ApJS 239.2, 18, p. 18. DOI: 10.3847/1538-4365/aae9f0.

Abdo, A. A., Ackermann, M., Ajello, M., et al. (May 2010). 'The First Catalog of Active Galactic Nuclei Detected by the Fermi Large Area Telescope'. In: ApJ 715.1, pp. 429–457. DOI: 10.1088/0004-637X/715/1/429.

Abramson, I. S. (1982). 'On Bandwidth Variation in Kernel Estimates-A Square Root Law'. In: *The Annals of Statistics* 10.4, pp. 1217–1223. ISSN: 00905364. URL: http://www.jstor.org/stable/2240724 (visited on 20/01/2024).

Adam, A., Perreault-Levasseur, L., Hezaveh, Y., and Welling, M. (July 2023). 'Pixelated Reconstruction of Foreground Density and Background Surface Brightness in Gravitational Lensing Systems Using Recurrent Inference Machines'. In: ApJ 951.1, 6, p. 6. DOI: 10.3847/1538-4357/accf84.

Aguado, D. S., Ahumada, R., Almeida, A., et al. (Feb. 2019). 'The Fifteenth Data Release of the Sloan Digital Sky Surveys: First Release of MaNGA-derived Quantities, Data Visualization Tools, and Stellar Library'. In: ApJS 240.2, 23, p. 23. DOI: 10.3847/1538-4365/aaf651.

Agüeros, M. A., Ivezić, Ž., Covey, K. R., et al. (Sept. 2005). 'The Ultraviolet, Optical, and Infrared Properties of Sloan Digital Sky Survey Sources Detected by GALEX'. In: AJ 130.3, pp. 1022–1036. DOI: 10.1086/432160.

Ahumada, R., Prieto, C. A., Almeida, A., et al. (July 2020). 'The 16th Data Release of the Sloan Digital Sky Surveys: First Release from the APOGEE-2 Southern Survey and Full Release of eBOSS Spectra'. In: ApJS 249.1, 3, p. 3. DOI: 10.3847/1538-4365/ab929e.

Aihara, H., Allende Prieto, C., An, D., et al. (Apr. 2011). 'The Eighth Data Release of the Sloan Digital Sky Survey: First Data from SDSS-III'. In: ApJS 193.2, 29, p. 29. DOI: 10.1088/0067-0049/193/2/29.

Alegre, L., Sabater, J., Best, P., et al. (Nov. 2022). 'A machine-learning classifier for LOFAR radio galaxy cross-matching techniques'. In: MNRAS 516.4, pp. 4716–4738. DOI: 10.1093/mnras/stac1888.

Ali, M. (Apr. 2020). *PyCaret: An open source, low-code machine learning library in Python*. PyCaret version 2.3. URL: https://www.pycaret.org.

Allen, D. M. (1974). 'The Relationship Between Variable Selection and Data Agumentation and a Method for Prediction'. In: *Technometrics* 16.1, pp. 125–127. DOI: 10.1080/00401706.1974.10489157.

Allison, P. (2001). *Missing Data*. Quantitative Applications in the Social Sciences. SAGE Publications. ISBN: 9781452207902. URL: https://books.google.pt/books?id=LJB2AwAAQBAJ.

Almosallam, I. A., Jarvis, M. J., and Roberts, S. J. (Oct. 2016a). 'GPZ: non-stationary sparse Gaussian processes for heteroscedastic uncertainty estimation in photometric redshifts'. In: MNRAS 462.1, pp. 726–739. DOI: 10.1093/mnras/stw1618.

Almosallam, I. A., Lindsay, S. N., Jarvis, M. J., and Roberts, S. J. (Jan. 2016b). 'A sparse Gaussian process framework for photometric redshift estimation'. In: MNRAS 455.3, pp. 2387–2401. DOI: 10.1093/mnras/stv2425.

Alqasim, A. and Page, M. J. (Apr. 2023). 'A new method to determine X-ray luminosity functions of AGN and their evolution with redshift'. In: MNRAS 520.3, pp. 3827–3846. DOI: 10.1093/mnras/stad007.

Amarantidis, S., Afonso, J., Messias, H., et al. (May 2019). 'The first supermassive black holes: indications from models for future observations'. In: MNRAS 485.2, pp. 2694–2709. DOI: 10.1093/mnras/stz551.

An, T., Zhang, Y., and Frey, S. (Sept. 2020). 'A method for checking high-redshift identification of radio AGNs'. In: MNRAS 497.2, pp. 2260–2264. DOI: 10.1093/mnras/staa2132.

Ananna, T. T., Salvato, M., LaMassa, S., et al. (Nov. 2017). 'AGN Populations in Large-volume X-Ray Surveys: Photometric Redshifts and Population Types Found in the Stripe 82X Survey'. In: ApJ 850.1, 66, p. 66. DOI: 10.3847/1538-4357/aa937d.

Anbajagane, D., Evrard, A. E., and Farahi, A. (Jan. 2022). 'Baryonic imprints on DM haloes: population statistics from dwarf galaxies to galaxy clusters'. In: MNRAS 509.3, pp. 3441–3461. DOI: 10.1093/mnras/stab3177.

Andonie, C., Alexander, D. M., Rosario, D., et al. (Dec. 2022). 'A panchromatic view of infrared quasars: excess star formation and radio emission in the most heavily obscured systems'. In: MNRAS 517.2, pp. 2577–2598. DOI: 10.1093/mnras/stac2800.

Aniyan, A. K. and Thorat, K. (June 2017). 'Classifying Radio Galaxies with the Convolutional Neural Network'. In: ApJS 230.2, 20, p. 20. DOI: 10.3847/1538-4365/aa7333.

Annis, J., Soares-Santos, M., Strauss, M. A., et al. (Oct. 2014). 'The Sloan Digital Sky Survey Coadd: 275 deg^2 of Deep Sloan Digital Sky Survey Imaging on Stripe 82'. In: ApJ 794.2, 120, p. 120. DOI: 10.1088/0004-637X/794/2/120.

Arévalo, P., Uttley, P., Kaspi, S., et al. (Sept. 2008). 'Correlated X-ray/optical variability in the quasar MR2251-178'. In: MNRAS 389.3, pp. 1479–1488. DOI: 10.1111/j.1365-2966.2008.13719.x.

Arévalo, P., Uttley, P., Lira, P., et al. (Aug. 2009). 'Correlation and time delays of the X-ray and optical emission of the Seyfert Galaxy NGC 3783'. In: MNRAS 397.4, pp. 2004–2014. DOI: 10.1111/j.1365-2966.2009.15110.x.

Arnouts, S., Cristiani, S., Moscardini, L., et al. (Dec. 1999). 'Measuring and modelling the redshift evolution of clustering: the Hubble Deep Field North'. In: MNRAS 310.2, pp. 540–556. DOI: 10.1046/j.1365-8711.1999.02978.x.

Arsioli, B. and Dedin, P. (Oct. 2020). 'Machine learning applied to multifrequency data in astrophysics: blazar classification'. In: MNRAS 498.2, pp. 1750–1764. DOI: 10.1093/mnras/staa2449.

Assef, R. J., Stern, D., Kochanek, C. S., et al. (July 2013). 'Mid-infrared Selection of Active Galactic Nuclei with the Wide-field Infrared Survey Explorer. II. Properties of WISE-selected Active Galactic Nuclei in the NDWFS Boötes Field'. In: ApJ 772.1, 26, p. 26. DOI: 10.1088/0004-637X/772/1/26.

Assef, R. J., Stern, D., Noirot, G., et al. (Feb. 2018). 'The WISE AGN Catalog'. In: ApJS 234.2, 23, p. 23. DOI: 10.3847/1538-4365/aaa00a.

Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. (Sept. 2018). 'The Astropy Project: Building an Open-science Project and Status of the v2.0 Core Package'. In: AJ 156.3, 123, p. 123. DOI: 10.3847/1538-3881/aabc4f.

Astropy Collaboration, Price-Whelan, A. M., Lim, P. L., et al. (Aug. 2022). 'The Astropy Project: Sustaining and Growing a Community-oriented Open-source Project and the Latest Major Release (v5.0) of the Core Package'. In: ApJ 935.2, 167, p. 167. DOI: 10.3847/1538-4357/ac7c74.

Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. (Oct. 2013). 'Astropy: A community Python package for astronomy'. In: A&A 558, A33, A33. DOI: 10.1051/0004-6361/201322068.

Atek, H., Chemerynska, I., Wang, B., et al. (Oct. 2023). 'JWST UNCOVER: discovery of z > 9 galaxy candidates behind the lensing cluster Abell 2744'. In: MNRAS 524.4, pp. 5486–5496. DOI: 10.1093/mnras/stad1998.

Auge, C., Sanders, D., Treister, E., et al. (Nov. 2023). 'The Accretion History of AGN: The Spectral Energy Distributions of X-Ray-luminous Active Galactic Nuclei'. In: ApJ 957.1, 19, p. 19. DOI: 10.3847/1538-4357/acf21a.

Bahcall, J. N. and Kozlovsky, B.-Z. (Mar. 1969). 'Some Models of the Emission-Line Region of 3c 273'. In: ApJ 155, p. 1077. DOI: 10.1086/149935.

Baldwin, J. A., Phillips, M. M., and Terlevich, R. (Feb. 1981). 'Classification parameters for the emission-line spectra of extragalactic objects.' In: PASP 93, pp. 5–19. DOI: 10.1086/130766.

Ball, N. M. and Brunner, R. J. (Jan. 2010). 'Data Mining and Machine Learning in Astronomy'. In: *International Journal of Modern Physics D* 19.7, pp. 1049–1106. DOI: 10.1142/S0218271810017160.

Ball, N. M., Brunner, R. J., Myers, A. D., et al. (Aug. 2008). 'Robust Machine Learning Applied to Astronomical Data Sets. III. Probabilistic Photometric Redshifts for Galaxies and Quasars in the SDSS and GALEX'. In: ApJ 683.1, pp. 12–21. DOI: 10.1086/589646.

Baltay, C., Grossman, L., Howard, R., et al. (Apr. 2021). 'Low-redshift Type Ia Supernova from the LSQ/LCO Collaboration'. In: PASP 133.1022, 044002, p. 044002. DOI: 10.1088/1538-3873/abd417.

Barbieri, C. and Bertola, F. (Jan. 1972). 'Identification of 5C4 radio sources.' In: MNRAS 156, pp. 399–409. DOI: 10.1093/mnras/156.4.399.

Baron, D. (Apr. 2019). 'Machine Learning in Astronomy: a practical overview'. In: *arXiv e-prints*, arXiv:1904.07248, arXiv:1904.07248.

Baron, D. and Poznanski, D. (Mar. 2017). 'The weirdest SDSS galaxies: results from an outlier detection algorithm'. In: MNRAS 465.4, pp. 4530–4555. DOI: 10.1093/mnras/stw3021.

Barrows, R. S., Comerford, J. M., Stern, D., and Assef, R. J. (Dec. 2021). 'A Catalog of Host Galaxies for WISE-selected AGN: Connecting Host Properties with Nuclear Activity and Identifying Contaminants'. In: ApJ 922.2, 179, p. 179. DOI: 10.3847/1538-4357/ac1352.

Baum, W. A. (Feb. 1957). 'Photoelectric determinations of redshifts beyond 0.2 c.' In: AJ 62, pp. 6–7. DOI: 10.1086/107433.

— (Jan. 1962). 'Photoelectric Magnitudes and Red-Shifts'. In: *Problems of Extra-Galactic Research*. Ed. by G. C. McVittie. Vol. 15, p. 390.

Beifiori, A., Courteau, S., Corsini, E. M., and Zhu, Y. (Jan. 2012). 'On the correlations between galaxy properties and supermassive black hole mass'. In: MNRAS 419.3, pp. 2497–2528. DOI: 10.1111/j.1365-2966.2011.19903.x.

Benítez, N. (June 2000). 'Bayesian Photometric Redshift Estimation'. In: ApJ 536.2, pp. 571–583. DOI: 10.1086/308947.

Bianchi, L., Rodriguez-Merino, L., Viton, M., et al. (Dec. 2007). 'Statistical Properties of the GALEX-SDSS Matched Source Catalogs, and Classification of the UV Sources'. In: ApJS 173.2, pp. 659–672. DOI: 10.1086/516648.

Birchall, K. L., Watson, M. G., and Aird, J. (Feb. 2020). 'X-ray detected AGN in SDSS dwarf galaxies'. In: MNRAS 492.2, pp. 2268–2284. DOI: 10.1093/mnras/staa040.

Blandford, R., Meier, D., and Readhead, A. (Aug. 2019). 'Relativistic Jets from Active Galactic Nuclei'. In: ARA&A 57, pp. 467–509. DOI: 10.1146/annurev-astro-081817-051948.

Blecha, L., Snyder, G. F., Satyapal, S., and Ellison, S. L. (Aug. 2018). 'The power of infrared AGN selection in mergers: a theoretical study'. In: MNRAS 478.3, pp. 3056–3071. DOI: 10.1093/mnras/sty1274.

Bonaldi, A., Bonato, M., Galluzzi, V., et al. (Jan. 2019). 'The Tiered Radio Extragalactic Continuum Simulation (T-RECS)'. In: MNRAS 482.1, pp. 2–19. DOI: 10.1093/mnras/sty2603.

Bonato, M., Prandoni, I., De Zotti, G., et al. (Jan. 2021). 'New constraints on the 1.4 GHz source number counts and luminosity functions in the Lockman Hole field'. In: MNRAS 500.1, pp. 22–33. DOI: 10.1093/mnras/staa3218.

Bonnarel, F., Fernique, P., Bienaymé, O., et al. (Apr. 2000). 'The ALADIN interactive sky atlas. A reference tool for identification of astronomical sources'. In: A&AS 143, pp. 33–40. DOI: 10.1051/aas:2000331.

Bouwens, R., González-López, J., Aravena, M., et al. (Oct. 2020). 'The ALMA Spectroscopic Survey Large Program: The Infrared Excess of z = 1.5-10 UV-selected Galaxies and the Implied High-redshift Star Formation History'. In: ApJ 902.2, 112, p. 112. DOI: 10.3847/1538-4357/abb830.

Bowler, R. A. A., Adams, N. J., Jarvis, M. J., and Häußler, B. (Mar. 2021). 'The rapid transition from star formation to AGN-dominated rest-frame ultraviolet light at z ≃ 4'. In: MNRAS 502.1, pp. 662–677. DOI: 10.1093/mnras/stab038.

Brammer, G. B., van Dokkum, P. G., and Coppi, P. (Oct. 2008). 'EAZY: A Fast, Public Photometric Redshift Code'. In: ApJ 686.2, pp. 1503–1513. DOI: 10.1086/591786.

Brandt, W. N. and Alexander, D. M. (Jan. 2015). 'Cosmic X-ray surveys of distant active galaxies. The demographics, physics, and ecology of growing supermassive black holes'. In: A&A Rev. 23, 1, p. 1. DOI: 10.1007/s00159-014-0081-z.

Braun, R., Bonaldi, A., Bourke, T., et al. (Dec. 2019). 'Anticipated Performance of the Square Kilometre Array – Phase 1 (SKA1)'. In: *arXiv e-prints*, arXiv:1912.12699, arXiv:1912.12699. DOI: 10.48550/arXiv.1912.12699.

Breedt, E., Arévalo, P., McHardy, I. M., et al. (Mar. 2009). 'Long-term optical and X-ray variability of the Seyfert galaxy Markarian 79'. In: MNRAS 394.1, pp. 427–437. DOI: 10.1111/j.1365-2966.2008.14302.x.

Breedt, E., McHardy, I. M., Arévalo, P., et al. (Apr. 2010). 'Twelve years of X-ray and optical variability in the Seyfert galaxy NGC 4051'. In: MNRAS 403.2, pp. 605–619. DOI: 10.1111/j.1365-2966.2009.16146.x.

Breiman, L. (Aug. 1996). 'Bagging predictors'. In: *Machine Learning* 24.2, pp. 123–140. ISSN: 1573-0565. DOI: 10.1007/BF00058655.

— (Oct. 2001). 'Random Forests'. In: *Machine Learning* 45.1, pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324.

— (2003). 'Manual on setting up, using, and understanding random forests v4. 0'. In: *Statistics Department University of California Berkeley, CA, USA*.

Brescia, M., Cavuoti, S., Razim, O., et al. (2021). 'Photometric Redshifts With Machine Learning, Lights and Shadows on a Complex Data Science Use Case'. In: *Frontiers in Astronomy and Space Sciences* 8, p. 70. ISSN: 2296-987X. DOI: 10.3389/fspas.2021.658229.

Brescia, M., Salvato, M., Cavuoti, S., et al. (Oct. 2019). 'Photometric redshifts for X-ray-selected active galactic nuclei in the eROSITA era'. In: MNRAS 489.1, pp. 663–680. DOI: 10.1093/mnras/stz2159.

Brier, G. W. (1950). 'Verification of Forecasts Expressed in Terms of Probability'. In: *Monthly Weather Review* 78.1, pp. 1–3. DOI: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Bröcker, J. and Smith, L. A. (2007). 'Increasing the Reliability of Reliability Diagrams'. In: *Weather and Forecasting* 22.3, pp. 651–661. DOI: 10.1175/WAF993.1.

Brown, M. J. I., Duncan, K. J., Landt, H., et al. (Nov. 2019). 'The spectral energy distributions of active galactic nuclei'. In: MNRAS 489.3, pp. 3351–3367. DOI: 10.1093/mnras/stz2324.

Brown, M. J. I., Moustakas, J., Smith, J. .-. T., et al. (June 2014). 'An Atlas of Galaxy Spectral Energy Distributions from the Ultraviolet to the Mid-infrared'. In: ApJS 212.2, 18, p. 18. DOI: 10.1088/0067-0049/212/2/18.

Brunner, H., Liu, T., Lamer, G., et al. (May 2022). 'The eROSITA Final Equatorial Depth Survey (eFEDS). X-ray catalogue'. In: A&A 661, A1, A1. DOI: 10.1051/0004-6361/202141266.

Brusa, M., Zamorani, G., Comastri, A., et al. (Sept. 2007). 'The XMM-Newton Wide-Field Survey in the COSMOS Field. III. Optical Identification and Multiwavelength Properties of a Large Sample of X-Ray-Selected Sources'. In: ApJS 172.1, pp. 353–367. DOI: 10.1086/516575.

Buchner, J. (Oct. 2019). 'Collaborative Nested Sampling: Big Data versus Complex Physical Models'. In: PASP 131.1004, p. 108005. DOI: 10.1088/1538-3873/aae7fc.

Budavári, T. and Szalay, A. S. (May 2008). 'Probabilistic Cross-Identification of Astronomical Sources'. In: ApJ 679.1, pp. 301–309. DOI: 10.1086/587156.

Buisson, D. J. K., Lohfink, A. M., Alston, W. N., and Fabian, A. C. (Jan. 2017). 'Ultraviolet and X-ray variability of active galactic nuclei with Swift'. In: MNRAS 464.3, pp. 3194–3218. DOI: 10.1093/mnras/stw2486.

Burhanudin, U. F., Maund, J. R., Killestein, T., et al. (Aug. 2021). 'Light-curve classification with recurrent neural networks for GOTO: dealing with imbalanced data'. In: MNRAS 505.3, pp. 4345–4361. DOI: 10.1093/mnras/stab1545.

Capetti, A., Brienza, M., Baldi, R. D., et al. (Oct. 2020). 'The LOFAR view of FR 0 radio galaxies'. In: A&A 642, A107, A107. DOI: 10.1051/0004-6361/202038671.

Carroll, B. W. and Ostlie, D. A. (2017). *An Introduction to Modern Astrophysics*. 2nd ed. Cambridge University Press. DOI: 10.1017/9781108380980.

Caruana, R. and Niculescu-Mizil, A. (2006). 'An Empirical Comparison of Supervised Learning Algorithms'. In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. Pittsburgh,

Pennsylvania, USA: Association for Computing Machinery, pp. 161–168. ISBN: 1595933832. DOI: 10.1145/1143844.1143865.

Carvajal, R., Bauer, F. E., Bouwens, R. J., et al. (Jan. 2020). 'The ALMA Frontier Fields Survey. V. ALMA Stacking of Lyman-Break Galaxies in Abell 2744, Abell 370, Abell S1063, MACSJ0416.1-2403 and MACSJ1149.5+2223'. In: A&A 633, A160, A160. DOI: 10.1051/0004-6361/201936260.

Carvajal, R., Matute, I., Afonso, J., et al. (Nov. 2023a). 'Selection of powerful radio galaxies with machine learning'. In: A&A 679, A101, A101. DOI: 10.1051/0004-6361/202245770.

Carvajal, R., Matute, I., Afonso, J., et al. (Oct. 2021). 'Exploring New Redshift Indicators for Radio-Powerful AGN'. In: *Galaxies* 9.4, p. 86. DOI: 10.3390/galaxies9040086.

Carvajal, R., Matute, I., Afonso, J., et al. (Dec. 2023b). *Selection of powerful radio galaxies with machine learning*. DOI: 10.5281/zenodo.10220009.

Casalicchio, G., Molnar, C., and Bischl, B. (2019). 'Visualizing the Feature Importance for Black Box Models'. In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by M. Berlingerio, F. Bonchi, T. Gärtner, et al. Cham: Springer International Publishing, pp. 655–670. ISBN: 978-3-030-10925-7.

Ceccarelli, L., Duplancic, F., and Garcia Lambas, D. (Jan. 2022). 'The impact of void environment on AGN'. In: MNRAS 509.2, pp. 1805–1819. DOI: 10.1093/mnras/stab2902.

Chambers, K. C., Magnier, E. A., Metcalfe, N., et al. (Dec. 2016). 'The Pan-STARRS1 Surveys'. In: *arXiv e-prints*, arXiv:1612.05560.

Champagne, J. B., Casey, C. M., Finkelstein, S. L., et al. (Aug. 2023). 'A Mixture of LBG Overdensities in the Fields of Three 6 < z < 7 Quasars: Implications for the Robustness of Photometric Selection'. In: ApJ 952.2, 99, p. 99. DOI: 10.3847/1538-4357/acda8d.

Chattopadhyay, A. K. (2017). 'Incomplete Data in Astrostatistics'. In: *Wiley StatsRef: Statistics Reference Online*. American Cancer Society, pp. 1–12. ISBN: 9781118445112. DOI: https://doi.org/10.1002/9781118445112.stat07942.

Chaves-Montero, J., Bonoli, S., Salvato, M., et al. (Dec. 2017). 'ELDAR, a new method to identify AGN in multi-filter surveys: the ALHAMBRA test case'. In: MNRAS 472.2, pp. 2085–2106. DOI: 10.1093/mnras/stx2054.

Chen, H., Garrett, M. A., Chi, S., et al. (June 2020). 'Searching for obscured AGN in z ∼ 2 submillimetre galaxies'. In: A&A 638, A113, A113. DOI: 10.1051/0004-6361/201937162.

Chen, T. and Guestrin, C. (2016). 'XGBoost: A Scalable Tree Boosting System'. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, pp. 785–794. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785.

Cid Fernandes, R., Stasińska, G., Mateus, A., and Vale Asari, N. (May 2011). 'A comprehensive classification of galaxies in the Sloan Digital Sky Survey: how to tell true from fake AGN?' In: MNRAS 413.3, pp. 1687–1699. DOI: 10.1111/j.1365-2966.2011.18244.x.

Cid Fernandes, R., Stasińska, G., Schlickmann, M. S., et al. (Apr. 2010). 'Alternative diagnostic diagrams and the 'forgotten' population of weak line galaxies in the SDSS'. In: MNRAS 403.2, pp. 1036–1053. DOI: 10.1111/j.1365-2966.2009.16185.x.

Clarke, A. O., Scaife, A. M. M., Greenhalgh, R., and Griguta, V. (July 2020). 'Identifying galaxies, quasars, and stars with machine learning: A new catalogue of classifications for 111 million SDSS sources without spectra'. In: A&A 639, A84, A84. DOI: 10.1051/0004-6361/201936770.

Cochrane, R. K., Kondapally, R., Best, P. N., et al. (Aug. 2023). 'The LOFAR Two-metre Sky Survey: the radio view of the cosmic star formation history'. In: MNRAS 523.4, pp. 6082–6102. DOI: 10.1093/mnras/stad1602.

Condon, J. J. (Jan. 1992). 'Radio emission from normal galaxies.' In: ARA&A 30, pp. 575–611. DOI: 10.1146/annurev.aa.30.090192.003043.

Condon, J. J., Cotton, W. D., and Broderick, J. J. (Aug. 2002). 'Radio Sources and Star Formation in the Local Universe'. In: AJ 124.2, pp. 675–689. DOI: 10.1086/341650.

Cortes, C. and Vapnik, V. (Sept. 1995). 'Support-vector networks'. In: *Machine Learning* 20.3, pp. 273–297. ISSN: 1573-0565. DOI: 10.1007/BF00994018.

Costa-Climent, R., Haftor, D. M., and Staniewski, M. W. (2023). 'Using machine learning to create and capture value in the business models of small and medium-sized enterprises'. In: *International Journal of Information Management* 73, p. 102637. ISSN: 0268-4012. DOI: https://doi.org/10.1016/j.ijinfomgt.2023.102637.

Cover, T. and Hart, P. (1967). 'Nearest neighbor pattern classification'. In: *IEEE Transactions on Information Theory* 13.1, pp. 21–27. DOI: 10.1109/TIT.1967.1053964.

Cramér, H. (1946). *Mathematical methods of statistics*. English. Princeton University Press Princeton, xvi, 575 p.

Cranmer, M. (May 2023). 'Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl'. In: *arXiv e-prints*, arXiv:2305.01582, arXiv:2305.01582. DOI: 10.48550/arXiv.2305.01582.

Cranmer, M., Sanchez-Gonzalez, A., Battaglia, P., et al. (June 2020). 'Discovering Symbolic Models from Deep Learning with Inductive Biases'. In: *arXiv e-prints*, arXiv:2006.11287.

Croom, S. M., Richards, G. T., Shanks, T., et al. (Nov. 2009). 'The 2dF-SDSS LRG and QSO survey: the QSO luminosity function at 0.4 < z < 2.6'. In: MNRAS 399.4, pp. 1755–1772. DOI: 10.1111/j.1365-2966.2009.15398.x.

Cunha, P. A. C. and Humphrey, A. (Oct. 2022). 'Photometric redshift-aided classification using ensemble learning'. In: A&A 666, A87, A87. DOI: 10.1051/0004-6361/202243135.

Curran, S. J. (May 2022). 'Quasar photometric redshifts from incomplete data using deep learning'. In: MNRAS 512.2, pp. 2099–2109. DOI: 10.1093/mnras/stac660.

Curran, S. J., Moss, J. P., and Perrott, Y. C. (July 2022). 'Redshifts of radio sources in the Million Quasars Catalogue from machine learning'. In: MNRAS 514.1, pp. 1–19. DOI: 10.1093/mnras/stac1333.

Cutri, R. M., Skrutskie, M. F., van Dyk, S., et al. (2003a). *2MASS All Sky Catalog of point sources.*

— (June 2003b). 'VizieR Online Data Catalog: 2MASS All-Sky Catalog of Point Sources (Cutri+ 2003)'. In: *VizieR Online Data Catalog*, II/246, pp. II/246.

Cutri, R. M., Wright, E. L., Conrow, T., et al. (Nov. 2013). *Explanatory Supplement to the AllWISE Data Release Products*.

Dahlen, T., Mobasher, B., Faber, S. M., et al. (Oct. 2013). 'A Critical Assessment of Photometric Redshift Methods: A CANDELS Investigation'. In: ApJ 775.2, 93, p. 93. DOI: 10.1088/0004-637X/775/2/93.

Davidson, K. and Netzer, H. (Oct. 1979). 'The emission lines of quasars and similar objects'. In: *Reviews of Modern Physics* 51.4, pp. 715–766. DOI: 10.1103/RevModPhys.51.715.

Davies, L. J. M., Robotham, A. S. G., Driver, S. P., et al. (Oct. 2018a). 'Deep Extragalactic VIsible Legacy Survey (DEVILS): motivation,design, and target catalogue'. In: MNRAS 480.1, pp. 768–799. DOI: 10.1093/mnras/sty1553.

Davies, T. M., Marshall, J. C., and Hazelton, M. L. (2018b). 'Tutorial on kernel estimation of continuous spatial and spatiotemporal relative risk'. In: *Statistics in Medicine* 37.7, pp. 1191–1221. DOI: https://doi.org/10.1002/sim.7577.

de Ruiter, H. R., Willis, A. G., and Arp, H. C. (May 1977). 'A Westerbork 1415 MHz survey of background radio sources. II. Optical identifications with deep IIIa-J plates.' In: A&AS 28, pp. 211–293.

Delhaize, J., Heywood, I., Prescott, M., et al. (Mar. 2021). 'MIGHTEE: are giant radio galaxies more common than we thought?' In: MNRAS 501.3, pp. 3833–3845. DOI: 10.1093/mnras/staa3837.

Delhaize, J., Smolčić, V., Delvecchio, I., et al. (June 2017). 'The VLA-COSMOS 3 GHz Large Project: The infrared-radio correlation of star-forming galaxies and AGN to $z \lesssim 6$'. In: A&A 602, A4, A4. DOI: 10.1051/0004-6361/201629430.

Desai, S. and Strachan, A. (June 2021). 'Parsimonious neural networks learn interpretable physical laws'. In: *Scientific Reports* 11.1, p. 12761. ISSN: 2045-2322. DOI: 10.1038/s41598-021-92278-w.

Dey, A., Schlegel, D. J., Lang, D., et al. (May 2019). 'Overview of the DESI Legacy Imaging Surveys'. In: AJ 157.5, 168, p. 168. DOI: 10.3847/1538-3881/ab089d.

Dey, B., Andrews, B. H., Newman, J. A., et al. (Oct. 2022). 'Photometric redshifts from SDSS images with an interpretable deep capsule network'. In: MNRAS 515.4, pp. 5285–5305. DOI: 10.1093/mnras/stac2105.

Dice, L. R. (1945). 'Measures of the Amount of Ecologic Association Between Species'. In: *Ecology* 26.3, pp. 297–302. ISSN: 00129658, 19399170. URL: http://www.jstor.org/stable/1932409 (visited on 04/10/2022).

Dobbels, W. and Baes, M. (Nov. 2021). 'Predicting far-infrared maps of galaxies via machine learning techniques'. In: A&A 655, A34, A34. DOI: 10.1051/0004-6361/202142084.

Donley, J. L., Koekemoer, A. M., Brusa, M., et al. (Apr. 2012). 'Identifying Luminous Active Galactic Nuclei in Deep Surveys: Revised IRAC Selection Criteria'. In: ApJ 748.2, 142, p. 142. DOI: 10.1088/0004-637X/748/2/142.

Dorogush, A. V., Ershov, V., and Gulin, A. (2018). 'CatBoost: gradient boosting with categorical features support'. In: *CoRR* abs/1810.11363. URL: http://arxiv.org/abs/1810.11363.

Drake, A. J., Graham, M. J., Djorgovski, S. G., et al. (July 2014). 'The Catalina Surveys Periodic Variable Star Catalog'. In: ApJS 213.1, 9, p. 9. DOI: 10.1088/0067-0049/213/1/9.

Driver, S. P., Hill, D. T., Kelvin, L. S., et al. (May 2011). 'Galaxy and Mass Assembly (GAMA): survey diagnostics and core data release'. In: MNRAS 413.2, pp. 971–995. DOI: 10.1111/j.1365-2966.2010.18188.x.

Duboue, P. (2020). *The Art of Feature Engineering: Essentials for Machine Learning*. Cambridge University Press. ISBN: 9781108709385. URL: https://books.google.pt/books?id=%5C_BzhDwAAQBAJ.

Duncan, K. J., Sabater, J., Röttgering, H. J. A., et al. (Feb. 2019). 'The LOFAR Two-metre Sky Survey. IV. First Data Release: Photometric redshifts and rest-frame magnitudes'. In: A&A 622, A3, A3. DOI: 10.1051/0004-6361/201833562.

Eddington, A. S. (Mar. 1913). 'On a formula for correcting statistics for the effects of a known error of observation'. In: MNRAS 73, pp. 359–360. DOI: 10.1093/mnras/73.5.359.

Eddington A. S., S. (Mar. 1940). 'The correction of statistics for accidental error'. In: MNRAS 100, p. 354. DOI: 10.1093/mnras/100.5.354.

Enke, H., Partl, A., Reinefeld, A., and Schintke, F. (Nov. 2012). 'Handling Big Data in Astronomy and Astrophysics: Rich Structured Queries on Replicated Cloud Data with XtreemFS'. In: *Datenbank-Spektrum* 12.3, pp. 173–181. ISSN: 1610-1995. DOI: 10.1007/s13222-012-0099-1.

Euclid Collaboration, Bisigello, L., Conselice, C. J., et al. (Apr. 2023a). 'Euclid preparation - XXIII. Derivation of galaxy physical properties with deep machine learning using mock fluxes and H-band images'. In: MNRAS 520.3, pp. 3529–3548. DOI: 10.1093/mnras/stac3810.

Euclid Collaboration, Humphrey, A., Bisigello, L., et al. (Mar. 2023b). 'Euclid preparation. XXII. Selection of quiescent galaxies from mock photometry using machine learning'. In: A&A 671, A99, A99. DOI: 10.1051/0004-6361/202244307.

Euclid Collaboration, Scaramella, R., Amiaux, J., et al. (June 2022). 'Euclid preparation. I. The Euclid Wide Survey'. In: A&A 662, A112, A112. DOI: 10.1051/0004-6361/202141938.

Fan, X., Banados, E., and Simcoe, R. A. (2023). 'Quasars and the Intergalactic Medium at Cosmic Dawn'. In: ARA&A 61. DOI: 10.1146/annurev-astro-052920-102455.

Ferrarese, L. and Merritt, D. (Aug. 2000). 'A Fundamental Relation between Supermassive Black Holes and Their Host Galaxies'. In: ApJ 539.1, pp. L9–L12. DOI: 10.1086/312838.

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

— (Dec. 2023). 'The Million Quasars (Milliquas) Catalogue, v8'. In: *The Open Journal of Astrophysics* 6, 49, p. 49. DOI: 10.21105/astro.2308.01505.

Flewelling, H. A., Magnier, E. A., Chambers, K. C., et al. (Nov. 2020). 'The Pan-STARRS1 Database and Data Products'. In: ApJS 251.1, 7, p. 7. DOI: 10.3847/1538-4365/abb82d.

Frederiksen, T. F., Graur, O., Hjorth, J., et al. (Mar. 2014). 'Spectroscopic identification of a redshift 1.55 supernova host galaxy from the Subaru Deep Field Supernova Survey'. In: A&A 563, A140, A140. DOI: 10.1051/0004-6361/201321795.

Freund, Y. and Schapire, R. E. (1996). 'Experiments with a New Boosting Algorithm'. In: *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*. ICML'96. Bari, Italy: Morgan Kaufmann Publishers Inc., pp. 148–156. ISBN: 1558604197.

Friedman, J. H. (2001). 'Greedy function approximation: A gradient boosting machine.' In: *The Annals of Statistics* 29.5, pp. 1189–1232. DOI: 10.1214/aos/1013203451.

— (2002). 'Stochastic gradient boosting'. In: *Computational Statistics & Data Analysis* 38.4. Nonlinear Methods and Data Mining, pp. 367–378. ISSN: 0167-9473. DOI: https://doi.org/10.1016/S0167-9473(01)00065-2.

Gaia Collaboration, Bailer-Jones, C. A. L., Teyssier, D., et al. (June 2023a). 'Gaia Data Release 3. The extragalactic content'. In: A&A 674, A41, A41. DOI: 10.1051/0004-6361/202243232.

Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., et al. (Nov. 2016). 'The Gaia mission'. In: A&A 595, A1, A1. DOI: 10.1051/0004-6361/201629272.

Gaia Collaboration, Vallenari, A., Brown, A. G. A., et al. (June 2023b). 'Gaia Data Release 3. Summary of the content and survey properties'. In: A&A 674, A1, A1. DOI: 10.1051/0004-6361/202243940.

Galametz, A., Grazian, A., Fontana, A., et al. (June 2013). 'CANDELS Multiwavelength Catalogs: Source Identification and Photometry in the CANDELS UKIDSS Ultra-deep Survey Field'. In: ApJS 206.2, 10, p. 10. DOI: 10.1088/0067-0049/206/2/10.

Garcia-Piquer, A., Morales, J. C., Ribas, I., et al. (Aug. 2017). 'Efficient scheduling of astronomical observations. Application to the CARMENES radial-velocity survey'. In: A&A 604, A87, A87. DOI: 10.1051/0004-6361/201628577.

Garilli, B., Fumana, M., Franzetti, P., et al. (July 2010). 'EZ: A Tool For Automatic Redshift Measurement'. In: PASP 122.893, p. 827. DOI: 10.1086/654903.

Garofalo, M., Botta, A., and Ventre, G. (June 2017). 'Astrophysics and Big Data: Challenges, Methods, and Tools'. In: *Astroinformatics*. Ed. by M. Brescia, S. G. Djorgovski, E. D. Feigelson, et al. Vol. 325, pp. 345–348. DOI: 10.1017/S1743921316012813.

Gebhardt, K., Bender, R., Bower, G., et al. (Aug. 2000). 'A Relationship between Nuclear Black Hole Mass and Galaxy Velocity Dispersion'. In: ApJ 539.1, pp. L13–L16. DOI: 10.1086/312840.

Gerwin, D. (1974). 'Information processing, data inferences, and scientific generalization'. In: *Behavioral Science* 19.5, pp. 314–325. DOI: https://doi.org/10.1002/bs.3830190504.

Getachew-Woreta, T., Pović, M., Masegosa, J., et al. (July 2022). 'Effect of AGN on the morphological properties of their host galaxies in the local Universe'. In: MNRAS 514.1, pp. 607–620. DOI: 10.1093/mnras/stac851.

Geurts, P., Ernst, D., and Wehenkel, L. (Apr. 2006). 'Extremely randomized trees'. In: *Machine Learning* 63.1, pp. 3–42. ISSN: 1573-0565. DOI: 10.1007/s10994-006-6226-1.

Giles, D. and Walkowicz, L. (Mar. 2019). 'Systematic serendipity: a test of unsupervised machine learning as a method for anomaly detection'. In: MNRAS 484.1, pp. 834–849. DOI: 10.1093/mnras/sty3461.

Giveon, U., Maoz, D., Kaspi, S., et al. (July 1999). 'Long-term optical variability properties of the Palomar-Green quasars'. In: MNRAS 306.3, pp. 637–654. DOI: 10.1046/j.1365-8711.1999.02556.x.

Glahn, H. R. and Jorgensen, D. L. (1970). 'Climatological Aspects of the Brier p-score'. In: *Monthly Weather Review* 98.2, pp. 136–141. DOI: 10.1175/1520-0493(1970)098<0136:CAOTBP>2.3.CO;2.

Glazebrook, K., Offer, A. R., and Deeley, K. (Jan. 1998). 'Automatic Redshift Determination by Use of Principal Component Analysis. I. Fundamentals'. In: ApJ 492.1, pp. 98–109. DOI: 10.1086/305039.

Glikman, E., Langgin, R., Johnstone, M. A., et al. (July 2023). 'A Candidate Dual QSO at Cosmic Noon'. In: ApJ 951.1, L18, p. L18. DOI: 10.3847/2041-8213/acda2f.

Goebel, R., Chander, A., Holzinger, K., et al. (2018). 'Explainable ai: the new 42?' In: *International cross-domain conference for machine learning and knowledge extraction*. Springer. Springer International Publishing, pp. 295–303. ISBN: 978-3-319-99740-7.

Gordon, Y. A., Boyce, M. M., O'Dea, C. P., et al. (Oct. 2020). 'A Catalog of Very Large Array Sky Survey Epoch 1 Quick Look Components, Sources, and Host Identifications'. In: *Research Notes of the American Astronomical Society* 4.10, 175, p. 175. DOI: 10.3847/2515-5172/abbe23.

Guglielmetti, F., Arras, P., Delli Veneri, M., et al. (Dec. 2022). 'Bayesian and Machine Learning Methods in the Big Data Era for Astronomical Imaging'. In: *Physical Sciences Forum*. Vol. 5. Physical Sciences Forum, 50, p. 50. DOI: 10.3390/psf2022005050.

Gültekin, K., Richstone, D. O., Gebhardt, K., et al. (June 2009). 'The M-σ and M-L Relations in Galactic Bulges, and Determinations of Their Intrinsic Scatter'. In: ApJ 698.1, pp. 198–221. DOI: 10.1088/0004-637X/698/1/198.

Habouzit, M., Onoue, M., Bañados, E., et al. (Apr. 2022). 'Co-evolution of massive black holes and their host galaxies at high redshift: discrepancies from six cosmological simulations and the key role of JWST'. In: MNRAS 511.3, pp. 3751–3767. DOI: 10.1093/mnras/stac225.

Hales, C. A., Murphy, T., Curran, J. R., et al. (Aug. 2012a). *BLOBCAT: Software to Catalog Blobs*. Astrophysics Source Code Library, record ascl:1208.009.

— (Sept. 2012b). 'BLOBCAT: software to catalogue flood-filled blobs in radio images of total intensity and linear polarization'. In: MNRAS 425.2, pp. 979–996. DOI: 10.1111/j.1365-2966.2012.21373.x.

Hancock, P. J., Murphy, T., Gaensler, B. M., et al. (May 2012). 'Compact continuum source finding for next generation radio surveys'. In: MNRAS 422.2, pp. 1812–1824. DOI: 10.1111/j.1365-2966.2012.20768.x.

Hancock, P. J., Trott, C. M., and Hurley-Walker, N. (Mar. 2018). 'Source Finding in the Era of the SKA (Precursors): Aegean 2.0'. In: PASA 35, e011, e011. DOI: 10.1017/pasa.2018.3.

Hardcastle, M. J., Horton, M. A., Williams, W. L., et al. (Oct. 2023). 'The LOFAR Two-Metre Sky Survey. VI. Optical identifications for the second data release'. In: A&A 678, A151, A151. DOI: 10.1051/0004-6361/202347333.

Häring, N. and Rix, H.-W. (Apr. 2004). 'On the Black Hole Mass-Bulge Mass Relation'. In: ApJ 604.2, pp. L89–L92. DOI: 10.1086/383567.

Head, T., Kumar, M., Nahrstaedt, H., et al. (Oct. 2021). *scikit-optimize/scikit-optimize*. Version v0.9.0. DOI: 10.5281/zenodo.5565057.

Heckman, T. M. and Best, P. N. (Aug. 2014). 'The Coevolution of Galaxies and Supermassive Black Holes: Insights from Surveys of the Contemporary Universe'. In: ARA&A 52, pp. 589–660. DOI: 10.1146/annurev-astro-081913-035722.

Helfand, D. J., White, R. L., and Becker, R. H. (Mar. 2015). 'The Last of FIRST: The Final Catalog and Source Identifications'. In: ApJ 801.1, 26, p. 26. DOI: 10.1088/0004-637X/801/1/26.

Helou, G., Soifer, B. T., and Rowan-Robinson, M. (Nov. 1985). 'Thermal infrared and nonthermal radio : remarkable correlation in disks of galaxies.' In: ApJ 298, pp. L7–L11. DOI: 10.1086/184556.

Hernán-Caballero, A., Varela, J., López-Sanjuan, C., et al. (Oct. 2021). 'The miniJPAS survey: Photometric redshift catalogue'. In: A&A 654, A101, A101. DOI: 10.1051/0004-6361/202141236.

Hickox, R. C. and Alexander, D. M. (Sept. 2018). 'Obscured Active Galactic Nuclei'. In: ARA&A 56, pp. 625–671. DOI: 10.1146/annurev-astro-081817-051803.

Hildebrand, R. H. (Sept. 1983). 'The determination of cloud masses and dust characteristics from submillimetre thermal emission.' In: QJRAS 24, pp. 267–282.

Hildebrandt, H., Arnouts, S., Capak, P., et al. (Nov. 2010). 'PHAT: PHoto-z Accuracy Testing'. In: A&A 523, A31, A31. DOI: 10.1051/0004-6361/201014885.

Hill, G. J., Gebhardt, K., Komatsu, E., et al. (Oct. 2008). 'The Hobby-Eberly Telescope Dark Energy Experiment (HETDEX): Description and Early Pilot Survey Results'. In: *Panoramic Views of Galaxy Formation and Evolution*. Ed. by T. Kodama, T. Yamada, and K. Aoki. Vol. 399. Astronomical Society of the Pacific Conference Series, p. 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. ISBN: 9780471097778. URL: https://books.google.pt/books?id=FRnvAAAAMAAJ.

Hodge, J. A., Becker, R. H., White, R. L., et al. (July 2011). 'High-resolution Very Large Array Imaging of Sloan Digital Sky Survey Stripe 82 at 1.4 GHz'. In: AJ 142.1, 3, p. 3. DOI: 10.1088/0004-6256/142/1/3.

Hogg, D. W. (May 1999). 'Distance measures in cosmology'. In: *arXiv e-prints*, astro-ph/9905116, astro–ph/9905116. DOI: 10.48550/arXiv.astro-ph/9905116.

Huertas-Company, M. and Lanusse, F. (Jan. 2023). 'The Dawes Review 10: The impact of deep learning for the analysis of galaxy surveys'. In: PASA 40, e001, e001. DOI: 10.1017/pasa.2022.55.

Hunter, J. D. (2007). 'Matplotlib: A 2D graphics environment'. In: *Computing in Science & Engineering* 9.3, pp. 90–95. DOI: 10.1109/MCSE.2007.55.

İkiz, T., Peletier, R. F., Barthel, P. D., and Yeşilyaprak, C. (Aug. 2020). 'Infrared-detected AGNs in the local Universe'. In: A&A 640, A68, A68. DOI: 10.1051/0004-6361/201935971.

Ilbert, O., Arnouts, S., McCracken, H. J., et al. (Oct. 2006). 'Accurate photometric redshifts for the CFHT legacy survey calibrated using the VIMOS VLT deep survey'. In: A&A 457.3, pp. 841–856. DOI: 10.1051/0004-6361:20065138.

Ilbert, O., Capak, P., Salvato, M., et al. (Jan. 2009). 'Cosmos Photometric Redshifts with 30-Bands for 2-deg^2'. In: ApJ 690.2, pp. 1236–1249. DOI: 10.1088/0004-637X/690/2/1236.

Inayoshi, K., Visbal, E., and Haiman, Z. (Aug. 2020). 'The Assembly of the First Massive Black Holes'. In: ARA&A 58, pp. 27–97. DOI: 10.1146/annurev-astro-120419-014455.

IPCC (2022). *Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by P. Shukla, J. Skea, R. Slade, et al. Cambridge, UK and New York, NY, USA: Cambridge University Press. DOI: 10.1017/9781009157926.

Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. (Mar. 2019). 'LSST: From Science Drivers to Reference Design and Anticipated Data Products'. In: ApJ 873.2, 111, p. 111. DOI: 10.3847/1538-4357/ab042c.

James, G., Witten, D., Hastie, T., et al. (2023). *An Introduction to Statistical Learning: with Applications in Python*. Springer Texts in Statistics. Springer International Publishing. ISBN: 9783031387470. URL: https://books.google.com/books?id=ygzJEAAAQBAJ.

Jarrett, T. H., Cluver, M. E., Magoulas, C., et al. (Feb. 2017). 'Galaxy and Mass Assembly (GAMA): Exploring the WISE Web in G12'. In: ApJ 836.2, 182, p. 182. DOI: 10.3847/1538-4357/836/2/182.

Jarvis, M., Taylor, R., Agudo, I., et al. (Jan. 2016). 'The MeerKAT International GHz Tiered Extragalactic Exploration (MIGHTEE) Survey'. In: *MeerKAT Science: On the Pathway to the SKA*, 6, p. 6. DOI: 10.22323/1.277.0006.

Jia, P., Jia, Q., Jiang, T., and Liu, J. (June 2023). 'Observation Strategy Optimization for Distributed Telescope Arrays with Deep Reinforcement Learning'. In: AJ 165.6, 233, p. 233. DOI: 10.3847/1538-3881/accceb.

Jiang, L., Fan, X., Bian, F., et al. (July 2014). 'The Sloan Digital Sky Survey Stripe 82 Imaging Data: Depth-optimized Co-adds over 300 deg^2 in Five Filters'. In: ApJS 213.1, 12, p. 12. DOI: 10.1088/0067-0049/213/1/12.

Jiang, T., Gradus, J. L., Lash, T. L., and Fox, M. P. (Feb. 2021). 'Addressing Measurement Error in Random Forests Using Quantitative Bias Analysis'. In: *American Journal of Epidemiology* 190.9, pp. 1830–1840. ISSN: 0002-9262. DOI: 10.1093/aje/kwab010.

Johnson, J., Douze, M., and Jégou, H. (2019). 'Billion-scale similarity search with GPUs'. In: *IEEE Transactions on Big Data* 7.3, pp. 535–547.

Johnson, N. and Leone, F. (1964). *Statistics and Experimental Design in Engineering and the Physical Sciences*. Vol. 2. Wiley, p. 125. ISBN: 9780471444893. URL: https://books.google.pt/books?id=IBjvAAAAMAAJ.

Jonas, J. and MeerKAT Team (Jan. 2016). 'The MeerKAT Radio Telescope'. In: *MeerKAT Science: On the Pathway to the SKA*, 1, p. 1. DOI: 10.22323/1.277.0001.

Josse, J., Prost, N., Scornet, E., and Varoquaux, G. (Feb. 2019). 'On the consistency of supervised learning with missing values'. In: *arXiv e-prints*, arXiv:1902.06931, arXiv:1902.06931. DOI: 10.48550/arXiv.1902.06931.

Josse, J. and Reiter, J. P. (2018). 'Introduction to the Special Section on Missing Data'. In: *Statistical Science* 33.2, pp. 139–141. DOI: 10.1214/18-STS332IN.

Kalton, G. and Kasprzyk, D. (1982). 'Imputing for missing survey responses'. In: *Proceedings of the section on survey research methods, American Statistical Association*. Vol. 22. American Statistical Association Cincinnati, p. 31.

Kauffmann, G., Heckman, T. M., Tremonti, C., et al. (Dec. 2003). 'The host galaxies of active galactic nuclei'. In: MNRAS 346.4, pp. 1055–1077. DOI: 10.1111/j.1365-2966.2003.07154.x.

Ke, G., Meng, Q., Finley, T., et al. (2017). 'LightGBM: A Highly Efficient Gradient Boosting Decision Tree'. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, et al. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

Kennicutt Robert C., J., Hao, C.-N., Calzetti, D., et al. (Oct. 2009). 'Dust-corrected Star Formation Rates of Galaxies. I. Combinations of Hα and Infrared Tracers'. In: ApJ 703.2, pp. 1672–1695. DOI: 10.1088/0004-637X/703/2/1672.

Kewley, L. J., Dopita, M. A., Sutherland, R. S., et al. (July 2001). 'Theoretical Modeling of Starburst Galaxies'. In: ApJ 556.1, pp. 121–140. DOI: 10.1086/321545.

Kewley, L. J., Groves, B., Kauffmann, G., and Heckman, T. (Nov. 2006). 'The host galaxies and classification of active galactic nuclei'. In: MNRAS 372.3, pp. 961–976. DOI: 10.1111/j.1365-2966.2006.10859.x.

Khramtsov, V., Spiniello, C., Agnello, A., and Sergeyev, A. (July 2021). 'VEXAS: VISTA EXtension to Auxiliary Surveys. Data Release 2: Machine-learning based classification of sources in the Southern Hemisphere'. In: A&A 651, A69, A69. DOI: 10.1051/0004-6361/202040131.

Kim, S. J., Lee, H. M., Matsuhara, H., et al. (Dec. 2012). 'The North Ecliptic Pole Wide survey of AKARI: a near- and mid-infrared source catalog'. In: A&A 548, A29, A29. DOI: 10.1051/0004-6361/201219105.

King, A. and Pounds, K. (Aug. 2015). 'Powerful Outflows and Feedback from Active Galactic Nuclei'. In: ARA&A 53, pp. 115–154. DOI: 10.1146/annurev-astro-082214-122316.

Kluyver, T., Ragan-Kelley, B., Pérez, F., et al. (2016). 'Jupyter Notebooks – a publishing format for reproducible computational workflows'. In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. Ed. by F. Loizides and B. Schmidt. IOS Press, pp. 87–90.

Kollmeier, J. A., Zasowski, G., Rix, H.-W., et al. (Nov. 2017). 'SDSS-V: Pioneering Panoptic Spectroscopy'. In: *arXiv e-prints*, arXiv:1711.03234, arXiv:1711.03234. DOI: 10.48550/arXiv.1711.03234.

Kondapally, R., Best, P. N., Cochrane, R. K., et al. (July 2022). 'Cosmic evolution of low-excitation radio galaxies in the LOFAR two-metre sky survey deep fields'. In: MNRAS 513.3, pp. 3742–3767. DOI: 10.1093/mnras/stac1128.

Kormendy, J. and Ho, L. C. (Aug. 2013). 'Coevolution (Or Not) of Supermassive Black Holes and Host Galaxies'. In: ARA&A 51.1, pp. 511–653. DOI: 10.1146/annurev-astro-082708-101811.

Koshida, S., Minezaki, T., Yoshii, Y., et al. (June 2014). 'Reverberation Measurements of the Inner Radius of the Dust Torus in 17 Seyfert Galaxies'. In: ApJ 788.2, 159, p. 159. DOI: 10.1088/0004-637X/788/2/159.

Koshida, S., Yoshii, Y., Kobayashi, Y., et al. (Aug. 2009). 'Variation of Inner Radius of Dust Torus in NGC4151'. In: ApJ 700.2, pp. L109–L113. DOI: 10.1088/0004-637X/700/2/L109.

Kovács, O. E., Bogdán, Á., Smith, R. K., et al. (Feb. 2019). 'Detection of the Missing Baryons toward the Sightline of H1821+643'. In: ApJ 872.1, 83, p. 83. DOI: 10.3847/1538-4357/aaef78.

Kull, M., Filho, T. M. S., and Flach, P. (2017a). 'Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration'. In: *Electronic Journal of Statistics* 11.2, pp. 5052–5080. DOI: 10.1214/17-EJS1338SI.

Kull, M., Filho, T. S., and Flach, P. (Apr. 2017b). 'Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers'. In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. Ed. by A. Singh and J. Zhu. Vol. 54. Proceedings of Machine Learning Research. PMLR, pp. 623–631. URL: https://proceedings.mlr.press/v54/kull17a.html.

Kurtz, M. J. and Mink, D. J. (Aug. 1998). 'RVSAO 2.0: Digital Redshifts and Radial Velocities'. In: PASP 110.750, pp. 934–977. DOI: 10.1086/316207.

Kuźmicz, A. and Jamrozy, M. (Mar. 2021). 'Giant Radio Quasars: Sample and Basic Properties'. In: ApJS 253.1, 25, p. 25. DOI: 10.3847/1538-4365/abd483.

Lacy, M., Baum, S. A., Chandler, C. J., et al. (Mar. 2020). 'The Karl G. Jansky Very Large Array Sky Survey (VLASS). Science Case and Survey Design'. In: PASP 132.1009, 035001, p. 035001. DOI: 10.1088/1538-3873/ab63eb.

Lacy, M., Ridgway, S. E., Gates, E. L., et al. (Oct. 2013). 'The Spitzer Mid-infrared Active Galactic Nucleus Survey. I. Optical and Near-infrared Spectroscopy of Obscured Candidates and Normal Active Galactic Nuclei Selected in the Mid-infrared'. In: ApJS 208.2, 24, p. 24. DOI: 10.1088/0067-0049/208/2/24.

Lacy, M., Storrie-Lombardi, L. J., Sajina, A., et al. (Sept. 2004). 'Obscured and Unobscured Active Galactic Nuclei in the Spitzer Space Telescope First Look Survey'. In: ApJS 154.1, pp. 166–169. DOI: 10.1086/422816.

Lacy, M., Surace, J. A., Farrah, D., et al. (Feb. 2021). 'A Spitzer survey of Deep Drilling Fields to be targeted by the Vera C. Rubin Observatory Legacy Survey of Space and Time'. In: MNRAS 501.1, pp. 892–910. DOI: 10.1093/mnras/staa3714.

Lacy, M. and Sajina, A. (Apr. 2020). 'Active galactic nuclei as seen by the Spitzer Space Telescope'. In: *Nature Astronomy* 4, pp. 352–363. DOI: 10.1038/s41550-020-1071-x.

Lal, D. V. (July 2021). 'The Discovery of a Remnant Radio Galaxy in A2065 Using GMRT'. In: ApJ 915.2, 126, p. 126. DOI: 10.3847/1538-4357/ac042d.

LaMassa, S. M., Urry, C. M., Cappelluti, N., et al. (Feb. 2016). 'The 31 Deg^2 Release of the Stripe 82 X-Ray Survey: The Point Source Catalog'. In: ApJ 817.2, 172, p. 172. DOI: 10.3847/0004-637X/817/2/172.

Lang, D. (May 2014). 'unWISE: Unblurred Coadds of the WISE Imaging'. In: AJ 147.5, 108, p. 108. DOI: 10.1088/0004-6256/147/5/108.

Langeroodi, D. and Hjorth, J. (Apr. 2023). 'PAH Emission from Star-forming Galaxies in JWST Mid-infrared Imaging of the Lensing Cluster SMACS J0723.3-7327'. In: ApJ 946.2, L40, p. L40. DOI: 10.3847/2041-8213/acc1e0.

Langley, P. (1977). 'BACON: A Production System That Discovers Empirical Laws'. In: *International Joint Conference on Artificial Intelligence*. URL: https://api.semanticscholar.org/CorpusID:2320342.

— (1979). 'Rediscovering Physics with BACON.3'. In: *Proceedings of the 6th International Joint Conference on Artificial Intelligence - Volume 1*. IJCAI'79. Tokyo, Japan: Morgan Kaufmann Publishers Inc., pp. 505–507. ISBN: 0934613478.

Langley, P., Bradshaw, G. L., and Simon, H. A. (1981). 'BAC0N.5: The Discovery of Conservation Laws'. In: *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 1*. IJCAI'81. Vancouver, BC, Canada: Morgan Kaufmann Publishers Inc., pp. 121–126.

Langley, P. and Zytkow, J. M. (1990). 'Data-Driven Approaches to Empirical Discovery'. In: *Machine Learning: Paradigms and Methods*. USA: Elsevier North-Holland, Inc., pp. 283–312. ISBN: 0262530880.

Latimer, C. J., Reines, A. E., Hainline, K. N., et al. (June 2021). 'A Chandra and HST View of WISE-selected AGN Candidates in Dwarf Galaxies'. In: ApJ 914.2, 133, p. 133. DOI: 10.3847/1538-4357/abfe0c.

Lazio, J. W., Kimball, A., Barger, A. J., et al. (Feb. 2014). 'Radio Astronomy in LSST Era'. In: PASP 126.936, p. 196. DOI: 10.1086/675262.

Le Fèvre, O., Tasca, L. A. M., Cassata, P., et al. (Apr. 2015). 'The VIMOS Ultra-Deep Survey: ~10 000 galaxies with spectroscopic redshifts to study galaxy assembly at early epochs $2 < z \simeq 6$'. In: A&A 576, A79, A79. DOI: 10.1051/0004-6361/201423829.

Lee, H. M., Kim, S. J., Im, M., et al. (Feb. 2009). 'North Ecliptic Pole Wide Field Survey of AKARI: Survey Strategy and Data Characteristics'. In: PASJ 61, p. 375. DOI: 10.1093/pasj/61.2.375.

Lehmer, B. D., Brandt, W. N., Alexander, D. M., et al. (Nov. 2005). 'The Extended Chandra Deep Field-South Survey: Chandra Point-Source Catalogs'. In: ApJS 161.1, pp. 21–40. DOI: 10.1086/444590.

Lichtenstein, S., Fischhoff, B., and Phillips, L. D. (1982). 'Calibration of probabilities: The state of the art to 1980'. In: *Judgment under Uncertainty: Heuristics and Biases*. Ed. by D. Kahneman, P. Slovic, and A. Tversky. Cambridge University Press, pp. 306–334. DOI: 10.1017/CBO9780511809477.023.

Lima, E. V. R., Sodré, L., Bom, C. R., et al. (Jan. 2022). 'Photometric redshifts for the S-PLUS Survey: Is machine learning up to the task?' In: *Astronomy and Computing* 38, 100510, p. 100510. DOI: 10.1016/j.ascom.2021.100510.

Lira, P., Arévalo, P., Uttley, P., et al. (Aug. 2011). 'Optical and near-IR long-term monitoring of NGC 3783 and MR 2251-178: evidence for variable near-IR emission from thin accretion discs'. In: MNRAS 415.2, pp. 1290–1303. DOI: 10.1111/j.1365-2966.2011.18774.x.

Lira, P., Arévalo, P., Uttley, P., et al. (Nov. 2015). 'Long-term monitoring of the archetype Seyfert galaxy MCG-6-30-15: X-ray, optical and near-IR variability of the corona, disc and torus'. In: MNRAS 454.1, pp. 368–379. DOI: 10.1093/mnras/stv1945.

Lisenfeld, U. and Völk, H. J. (Feb. 2000). 'On the radio spectral index of galaxies'. In: A&A 354, pp. 423–430.

Liske, J., Baldry, I. K., Driver, S. P., et al. (Sept. 2015). 'Galaxy And Mass Assembly (GAMA): end of survey report and data release 2'. In: MNRAS 452.2, pp. 2087–2126. DOI: 10.1093/mnras/stv1436.

Little, R. and Rubin, D. (2014). *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Wiley. ISBN: 9781118625880. URL: https://books.google.pt/books?id=AyVeBAAAQBAJ.

Liu, D., Lang, P., Magnelli, B., et al. (Oct. 2019). 'Automated Mining of the ALMA Archive in the COSMOS Field (A^3COSMOS). I. Robust ALMA Continuum Photometry Catalogs and Stellar Mass and Star Formation Properties for ~700 Galaxies at z = 0.5-6'. In: ApJS 244.2, 40, p. 40. DOI: 10.3847/1538-4365/ab42da.

Lochner, M. and Bassett, B. A. (July 2021). 'ASTRONOMALY: Personalised active anomaly detection in astronomical data'. In: *Astronomy and Computing* 36, 100481, p. 100481. DOI: 10.1016/j.ascom.2021.100481.

Loh, E. D. and Spillar, E. J. (Apr. 1986). 'Photometric Redshifts of Galaxies'. In: ApJ 303, p. 154. DOI: 10.1086/164062.

Louppe, G., Wehenkel, L., Sutera, A., and Geurts, P. (2013). 'Understanding variable importances in forests of randomized trees'. In: *Advances in Neural Information Processing Systems*. Ed. by C. J. C. Burges, L. Bottou, M. Welling, et al. Vol. 26. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2013/file/e3796ae838835da0b6f6ea37bcf8bcb7-Paper.pdf.

LSST Science Collaboration, Abell, P. A., Allison, J., et al. (Dec. 2009). 'LSST Science Book, Version 2.0'. In: *arXiv e-prints*, arXiv:0912.0201, arXiv:0912.0201. DOI: 10.48550/arXiv.0912.0201.

Luken, K., Norris, R., Park, L., et al. (2022). 'Estimating galaxy redshift in radio-selected datasets using machine learning'. In: *Astronomy and Computing* 39, p. 100557. ISSN: 2213-1337. DOI: https://doi.org/10.1016/j.ascom.2022.100557.

Luken, K. J., Norris, R. P., and Park, L. A. F. (Oct. 2019). 'Preliminary Results of Using k-Nearest Neighbors Regression to Estimate the Redshift of Radio-selected Data Sets'. In: PASP 131.1004, p. 108003. DOI: 10.1088/1538-3873/aaea17.

Lukic, V., Brüggen, M., Mingo, B., et al. (Aug. 2019). 'Morphological classification of radio galaxies: capsule networks versus convolutional neural networks'. In: MNRAS 487.2, pp. 1729–1744. DOI: 10.1093/mnras/stz1289.

Lundberg, S. M. and Lee, S.-I. (2017). 'A Unified Approach to Interpreting Model Predictions'. In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, et al. Curran Associates, Inc., pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Lundberg, S. M., Erion, G., Chen, H., et al. (2020). 'From local explanations to global understanding with explainable AI for trees'. In: *Nature Machine Intelligence* 2.1, pp. 2522–5839.

Lyu, J., Alberts, S., Rieke, G. H., and Rujopakarn, W. (Dec. 2022). 'AGN Selection and Demographics in GOODS-S/HUDF from X-Ray to Radio'. In: ApJ 941.2, 191, p. 191. DOI: 10.3847/1538-4357/ac9e5d.

Ma, S. and Tourani, R. (Aug. 2020). 'Predictive and Causal Implications of using Shapley Value for Model Interpretation'. In: *Proceedings of the 2020 KDD Workshop on Causal Discovery*. Vol. 127. Proceedings of Machine Learning Research. PMLR, pp. 23–38. URL: https://proceedings.mlr.press/v127/ma20a.html.

Ma, Z., Xu, H., Zhu, J., et al. (Feb. 2019). 'A Machine Learning Based Morphological Classification of 14,245 Radio AGNs Selected from the Best-Heckman Sample'. In: ApJS 240.2, 34, p. 34. DOI: 10.3847/1538-4365/aaf9a2.

Machado, D. P., Leonard, A., Starck, J. .-., et al. (Dec. 2013). 'Darth Fader: Using wavelets to obtain accurate redshifts of spectra at very low signal-to-noise'. In: A&A 560, A83, A83. DOI: 10.1051/0004-6361/201219857.

Machado Poletti Valle, L. F., Avestruz, C., Barnes, D. J., et al. (Oct. 2021). 'SHAPing the gas: understanding gas shapes in dark matter haloes with interpretable machine learning'. In: MNRAS 507.1, pp. 1468–1484. DOI: 10.1093/mnras/stab2252.

Madau, P. and Dickinson, M. (Aug. 2014). 'Cosmic Star-Formation History'. In: ARA&A 52, pp. 415–486. DOI: 10.1146/annurev-astro-081811-125615.

Magliocchetti, M., Lutz, D., Rosario, D., et al. (July 2014). 'The PEP survey: infrared properties of radio-selected AGN'. In: MNRAS 442.1, pp. 682–693. DOI: 10.1093/mnras/stu863.

Magliocchetti, M. (Dec. 2022). 'Hosts and environments: a (large-scale) radio history of AGN and star-forming galaxies'. In: A&A Rev. 30.1, 6, p. 6. DOI: 10.1007/s00159-022-00142-1.

Magliocchetti, M., Maddox, S. J., Jackson, C. A., et al. (June 2002). 'The 2dF Galaxy Redshift Survey: the population of nearby radio galaxies at the 1-mJy level'. In: MNRAS 333.1, pp. 100–120. DOI: 10.1046/j.1365-8711.2002.05386.x.

Magorrian, J., Tremaine, S., Richstone, D., et al. (June 1998). 'The Demography of Massive Dark Objects in Galaxy Centers'. In: AJ 115.6, pp. 2285–2305. DOI: 10.1086/300353.

Mainzer, A., Bauer, J., Cutri, R. M., et al. (Sept. 2014). 'Initial Performance of the NEOWISE Reactivation Mission'. In: ApJ 792.1, 30, p. 30. DOI: 10.1088/0004-637X/792/1/30.

Mainzer, A., Bauer, J., Grav, T., et al. (Apr. 2011). 'Preliminary Results from NEOWISE: An Enhancement to the Wide-field Infrared Survey Explorer for Solar System Science'. In: ApJ 731.1, 53, p. 53. DOI: 10.1088/0004-637X/731/1/53.

Maitra, C., Haberl, F., Ivanov, V. D., et al. (Feb. 2019). 'Identification of AGN in the XMM-Newton X-ray survey of the SMC'. In: A&A 622, A29, A29. DOI: 10.1051/0004-6361/201833663.

Mandal, S., Prandoni, I., Hardcastle, M. J., et al. (Apr. 2021). 'Extremely deep 150 MHz source counts from the LoTSS Deep Fields'. In: A&A 648, A5, A5. DOI: 10.1051/0004-6361/202039998.

Marchesi, S., Civano, F., Elvis, M., et al. (Jan. 2016). 'The Chandra COSMOS Legacy survey: optical/IR identifications'. In: ApJ 817.1, 34, p. 34. DOI: 10.3847/0004-637X/817/1/34.

Marocco, F., Eisenhardt, P. R. M., Fowler, J. W., et al. (Mar. 2021). 'The CatWISE2020 Catalog'. In: ApJS 253.1, 8, p. 8. DOI: 10.3847/1538-4365/abd805.

Mateos, S., Alonso-Herrero, A., Carrera, F. J., et al. (Nov. 2012). 'Using the Bright Ultrahard XMM-Newton survey to define an IR selection of luminous AGN based on WISE colours'. In: MNRAS 426.4, pp. 3271–3281. DOI: 10.1111/j.1365-2966.2012.21843.x.

Matthews, B. (1975). 'Comparison of the predicted and observed secondary structure of T4 phage lysozyme'. In: *Biochimica et Biophysica Acta (BBA) - Protein Structure* 405.2, pp. 442–451. ISSN: 0005-2795. DOI: https://doi.org/10.1016/0005-2795(75)90109-9.

Mauch, T. and Sadler, E. M. (Mar. 2007). 'Radio sources in the 6dFGS: local luminosity functions at 1.4 GHz for star-forming galaxies and radio-loud AGN'. In: MNRAS 375.3, pp. 931–950. DOI: 10.1111/j.1365-2966.2006.11353.x.

McAlpine, K., Jarvis, M. J., and Bonfield, D. G. (Dec. 2013). 'Evolution of faint radio sources in the VIDEO-XMM3 field'. In: MNRAS 436.2, pp. 1084–1095. DOI: 10.1093/mnras/stt1638.

McConnell, D., Hale, C. L., Lenc, E., et al. (Nov. 2020). 'The Rapid ASKAP Continuum Survey I: Design and first results'. In: PASA 37, e048, e048. DOI: 10.1017/pasa.2020.41.

McConnell, N. J. and Ma, C.-P. (Feb. 2013). 'Revisiting the Scaling Relations of Black Hole Masses and Host Galaxy Properties'. In: ApJ 764.2, 184, p. 184. DOI: 10.1088/0004-637X/764/2/184.

McGreer, I. D., Becker, R. H., Helfand, D. J., and White, R. L. (Nov. 2006). 'Discovery of a z = 6.1 Radio-Loud Quasar in the NOAO Deep Wide Field Survey'. In: ApJ 652.1, pp. 157–162. DOI: 10.1086/507767.

McHardy, I. M., Connolly, S. D., Peterson, B. M., et al. (May 2016). 'The origin of UV-optical variability in AGN and test of disc models: XMM-Newton and ground-based observations of NGC 4395'. In: *Astronomische Nachrichten* 337.4-5, p. 500. DOI: 10.1002/asna.201612337.

McKinney, W. (2010). 'Data Structures for Statistical Computing in Python'. In: *Proceedings of the 9th Python in Science Conference*. Ed. by S. van der Walt and J. Millman, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a.

Meisner, A. M., Lang, D., Schlafly, E. F., and Schlegel, D. J. (Dec. 2019). 'unWISE Coadds: The Five-year Data Set'. In: PASP 131.1006, p. 124504. DOI: 10.1088/1538-3873/ab3df4.

Menzel, M. .-., Merloni, A., Georgakakis, A., et al. (Mar. 2016). 'A spectroscopic survey of X-ray-selected AGNs in the northern XMM-XXL field'. In: MNRAS 457.1, pp. 110–132. DOI: 10.1093/mnras/stv2749.

Merlin, E., Castellano, M., Santini, P., et al. (May 2021). 'The ASTRODEEP-GS43 catalogue: New photometry and redshifts for the CANDELS GOODS-South field'. In: A&A 649, A22, A22. DOI: 10.1051/0004-6361/202140310.

Messias, H., Afonso, J., Salvato, M., et al. (Aug. 2012). 'A New Infrared Color Criterion for the Selection of 0 < z < 7 AGNs: Application to Deep Fields and Implications for JWST Surveys'. In: ApJ 754.2, 120, p. 120. DOI: 10.1088/0004-637X/754/2/120.

Michailidis, M. (2017). 'Investigating machine learning methods in recommender systems'. PhD thesis. University College London, UK. URL: https://discovery.ucl.ac.uk/id/eprint/10031000.

Michelucci, U. and Venturini, F. (2023). 'New metric formulas that include measurement errors in machine learning for natural sciences'. In: *Expert Systems with Applications* 224, p. 120013. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2023.120013.

Mickaelian, A. M. (Dec. 2020). 'Big Data in Astronomy: Surveys, Catalogs, Databases and Archives'. In: *Communications of the Byurakan Astrophysical Observatory* 67, pp. 159–180. DOI: 10.52526/25792776-2020.67.2-159.

Miley, G. and De Breuck, C. (Feb. 2008). 'Distant radio galaxies and their environments'. In: A&A Rev. 15.2, pp. 67–144. DOI: 10.1007/s00159-007-0008-z.

Mingo, B., Watson, M. G., Rosen, S. R., et al. (Nov. 2016). 'The MIXR sample: AGN activity versus star formation across the cross-correlation of WISE, 3XMM, and FIRST/NVSS'. In: MNRAS 462.3, pp. 2631–2667. DOI: 10.1093/mnras/stw1826.

Miyaji, T., Hasinger, G., and Schmidt, M. (Apr. 2001). 'Soft X-ray AGN luminosity function from ROSAT surveys. II. Table of the binned soft X-ray luminosity function'. In: A&A 369, pp. 49–56. DOI: 10.1051/0004-6361:20010102.

Mo, W., Gonzalez, A., Brodwin, M., et al. (Oct. 2020). 'The Massive and Distant Clusters of WISE Survey. VIII. Radio Activity in Massive Galaxy Clusters'. In: ApJ 901.2, 131, p. 131. DOI: 10.3847/1538-4357/abb08d.

Mohan, N. and Rafferty, D. (Feb. 2015). *PyBDSF: Python Blob Detection and Source Finder*. Astrophysics Source Code Library, record ascl:1502.007.

Morrissey, P., Conrow, T., Barlow, T. A., et al. (Dec. 2007). 'The Calibration and Data Products of GALEX'. In: ApJS 173.2, pp. 682–697. DOI: 10.1086/520512.

Mostert, R. I. J., Duncan, K. J., Röttgering, H. J. A., et al. (Jan. 2021). 'Unveiling the rarest morphologies of the LOFAR Two-metre Sky Survey radio source population with self-organised maps'. In: A&A 645, A89, A89. DOI: 10.1051/0004-6361/202038500.

Moya, A. and López-Sastre, R. J. (July 2022). 'Stellar mass and radius estimation using artificial intelligence'. In: A&A 663, A112, A112. DOI: 10.1051/0004-6361/202142930.

Naidoo, K., Johnston, H., Joachimi, B., et al. (Feb. 2023). 'Euclid: Calibrating photometric redshifts with spectroscopic cross-correlations'. In: A&A 670, A149, A149. DOI: 10.1051/0004-6361/202244795.

Nakoneczny, S. J., Bilicki, M., Pollo, A., et al. (May 2021). 'Photometric selection and redshifts for quasars in the Kilo-Degree Survey Data Release 4'. In: A&A 649, A81, A81. DOI: 10.1051/0004-6361/202039684.

Netzer, H. (Aug. 2015). 'Revisiting the Unified Model of Active Galactic Nuclei'. In: ARA&A 53, pp. 365–408. DOI: 10.1146/annurev-astro-082214-122302.

Newman, J. A., Abate, A., Abdalla, F. B., et al. (Mar. 2015). 'Spectroscopic needs for imaging dark energy experiments'. In: *Astroparticle Physics* 63, pp. 81–100. DOI: 10.1016/j.astropartphys.2014.06.007.

Newman, J. A. and Gruen, D. (Aug. 2022). 'Photometric Redshifts for Next-Generation Surveys'. In: ARA&A 60, pp. 363–414. DOI: 10.1146/annurev-astro-032122-014611.

Nicastro, F., Kaastra, J., Krongold, Y., et al. (June 2018). 'Observations of the missing baryons in the warm-hot intergalactic medium'. In: Nature 558.7710, pp. 406–409. DOI: 10.1038/s41586-018-0204-1.

Nicastro, F., Krongold, Y., Mathur, S., and Elvis, M. (Mar. 2017). 'A decade of warm hot intergalactic medium searches: Where do we stand and where do we go?' In: *Astronomische Nachrichten* 338.281, pp. 281–286. DOI: 10.1002/asna.201713343.

Niculescu-Mizil, A. and Caruana, R. (2005). 'Predicting Good Probabilities with Supervised Learning'. In: *Proceedings of the 22nd International Conference on Machine Learning*. ICML '05. Bonn, Germany: Association for Computing Machinery, pp. 625–632. ISBN: 1595931805. DOI: 10.1145/1102351.1102430.

Norris, R. P. (June 2017). 'Astroinformatics Challenges from Next-generation Radio Continuum Surveys'. In: *Astroinformatics*. Ed. by M. Brescia, S. G. Djorgovski, E. D. Feigelson, et al. Vol. 325, pp. 103–113. DOI: 10.1017/S1743921316012825.

Norris, R. P., Hopkins, A. M., Afonso, J., et al. (Aug. 2011). 'EMU: Evolutionary Map of the Universe'. In: PASA 28.3, pp. 215–248. DOI: 10.1071/AS11021.

Norris, R. P., Marvil, J., Collier, J. D., et al. (Sept. 2021). 'The Evolutionary Map of the Universe pilot survey'. In: PASA 38, e046, e046. DOI: 10.1017/pasa.2021.42.

Norris, R. P., Salvato, M., Longo, G., et al. (Oct. 2019). 'A Comparison of Photometric Redshift Techniques for Large Radio Surveys'. In: PASP 131.1004, p. 108004. DOI: 10.1088/1538-3873/ab0f7b.

Nour, D. and Sriram, K. (Jan. 2023). 'Association of optical, ultraviolet, and soft X-ray excess emissions in AGNs'. In: MNRAS 518.4, pp. 5705–5717. DOI: 10.1093/mnras/stac3505.

Ochsenbein, F., Bauer, P., and Marcout, J. (Apr. 2000). 'The VizieR database of astronomical catalogues'. In: A&AS 143, pp. 23–32. DOI: 10.1051/aas:2000169.

Oke, J. B. and Sandage, A. (Oct. 1968). 'Energy Distributions, K Corrections, and the Stebbins-Whitford Effect for Giant Elliptical Galaxies'. In: ApJ 154, p. 21. DOI: 10.1086/149737.

Oliver, S., Rowan-Robinson, M., Alexander, D. M., et al. (Aug. 2000). 'The European Large Area ISO Survey - I. Goals, definition and observations'. In: MNRAS 316.4, pp. 749–767. DOI: 10.1046/j.1365-8711.2000.03550.x.

Opitz, D. and Maclin, R. (July 1999). 'Popular Ensemble Methods: An Empirical Study'. In: *J. Artif. Int. Res.* 11.1, pp. 169–198. ISSN: 1076-9757.

Osorio-Clavijo, N., Gonzalez-Martín, O., Sánchez, S. F., et al. (July 2023). 'AGNs in the CALIFA survey: X-ray detection of nuclear sources'. In: MNRAS 522.4, pp. 5788–5804. DOI: 10.1093/mnras/stad1262.

Pacifici, C., Iyer, K. G., Mobasher, B., et al. (Feb. 2023). 'The Art of Measuring Physical Parameters in Galaxies: A Critical Assessment of Spectral Energy Distribution Fitting Techniques'. In: ApJ 944.2, 141, p. 141. DOI: 10.3847/1538-4357/acacff.

Padovani, P., Alexander, D. M., Assef, R. J., et al. (Aug. 2017). 'Active galactic nuclei: what's in a name?' In: A&A Rev. 25.1, 2, p. 2. DOI: 10.1007/s00159-017-0102-9.

Padovani, P. (Sept. 2016). 'The faint radio sky: radio astronomy becomes mainstream'. In: A&A Rev. 24.1, 13, p. 13. DOI: 10.1007/s00159-016-0098-6.

— (Nov. 2017). 'Active Galactic Nuclei at all wavelengths and from all angles'. In: *Frontiers in Astronomy and Space Sciences* 4, 35, p. 35. DOI: 10.3389/fspas.2017.00035.

Page, M. J. and Carrera, F. J. (Jan. 2000). 'An improved method of constructing binned luminosity functions'. In: MNRAS 311.2, pp. 433–440. DOI: 10.1046/j.1365-8711.2000.03105.x.

Palanque-Delabrouille, N., Magneville, C., Yèche, C., et al. (Mar. 2016). 'The extended Baryon Oscillation Spectroscopic Survey: Variability selection and quasar luminosity function'. In: A&A 587, A41, A41. DOI: 10.1051/0004-6361/201527392.

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). 'Scikit-learn: Machine Learning in Python'. In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Pepinsky, T. B. (2018). 'A Note on Listwise Deletion versus Multiple Imputation'. In: *Political Analysis* 26.4, pp. 480–488. DOI: 10.1017/pan.2018.18.

Pérez-Torres, M., Mattila, S., Alonso-Herrero, A., et al. (Dec. 2021). 'Star formation and nuclear activity in luminous infrared galaxies: an infrared through radio review'. In: A&A Rev. 29.1, 2, p. 2. DOI: 10.1007/s00159-020-00128-x.

Planck Collaboration, Aghanim, N., Akrami, Y., et al. (Sept. 2020). 'Planck 2018 results. VI. Cosmological parameters'. In: A&A 641, A6, A6. DOI: 10.1051/0004-6361/201833910.

Poisot, T. (2023). 'Guidelines for the prediction of species interactions through binary classification'. In: *Methods in Ecology and Evolution* 14.5, pp. 1333–1345. DOI: https://doi.org/10.1111/2041-210X.14071.

Poliszczuk, A., Pollo, A., Małek, K., et al. (July 2021). 'Active galactic nuclei catalog from the AKARI NEP-Wide field'. In: A&A 651, A108, A108. DOI: 10.1051/0004-6361/202040219.

Portegies Zwart, S. (Sept. 2020). 'The ecological impact of high-performance computing in astrophysics'. In: *Nature Astronomy* 4, pp. 819–822. DOI: 10.1038/s41550-020-1208-y.

Pouliasis, E. (Feb. 2020). 'Identification of Active Galactic Nuclei through different selection techniques'. PhD thesis. IAASARS, National Observatory of Athens.

Prandoni, I., Gregorini, L., Parma, P., et al. (Jan. 2001). 'The ATESP radio survey. III. Source counts'. In: A&A 365, pp. 392–399. DOI: 10.1051/0004-6361:20000142.

Prandoni, I., Guglielmino, G., Morganti, R., et al. (Dec. 2018). 'The Lockman Hole Project: new constraints on the sub-mJy source counts from a wide-area 1.4 GHz mosaic'. In: MNRAS 481.4, pp. 4548–4565. DOI: 10.1093/mnras/sty2521.

Prandoni, I. and Seymour, N. (Apr. 2015). 'Revealing the Physics and Evolution of Galaxies and Galaxy Clusters with SKA Continuum Surveys'. In: *Advancing Astrophysics with the Square Kilometre Array (AASKA14)*, 67, p. 67.

Predehl, P., Andritschke, R., Arefiev, V., et al. (Mar. 2021). 'The eROSITA X-ray telescope on SRG'. In: A&A 647, A1, A1. DOI: 10.1051/0004-6361/202039313.

Prestage, R. M. and Peacock, J. A. (July 1983). 'Optical identifications of Parkes radio sources using UK Schmidt plates.' In: MNRAS 204, pp. 355–364. DOI: 10.1093/mnras/204.2.355.

Prokhorenkova, L., Gusev, G., Vorobev, A., et al. (2018). 'CatBoost: unbiased boosting with categorical features'. In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, et al.

Vol. 31. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf.

Radcliffe, J. F., Barthel, P. D., Garrett, M. A., et al. (May 2021). 'The radio emission from active galactic nuclei'. In: A&A 649, L9, p. L9. DOI: 10.1051/0004-6361/202140791.

Radcliffe, J. F., Garrett, M. A., Muxlow, T. W. B., et al. (Nov. 2018). 'Nowhere to Hide: Radio-faint AGN in GOODS-N field. I. Initial catalogue and radio properties'. In: A&A 619, A48, A48. DOI: 10.1051/0004-6361/201833399.

Rajagopal, M., Marchesi, S., Kaur, A., et al. (June 2021). 'Identifying the 3FHL Catalog. V. Results of the CTIO-COSMOS Optical Spectroscopy Campaign 2019'. In: ApJS 254.2, 26, p. 26. DOI: 10.3847/1538-4365/abf656.

Rasmussen, C. E. and Williams, C. K. I. (Nov. 2005). *Gaussian Processes for Machine Learning*. The MIT Press. ISBN: 9780262256834. DOI: 10.7551/mitpress/3206.001.0001.

Ratner, B. (June 2009). 'The correlation coefficient: Its values range between +1/-1, or do they?' In: *Journal of Targeting, Measurement and Analysis for Marketing* 17.2, pp. 139–142. ISSN: 1479-1862. DOI: 10.1057/jt.2009.5.

Rawlings, S. (Sept. 2003). 'High-redshift radio galaxies: at the crossroads'. In: New A Rev. 47.4-5, pp. 397–404. DOI: 10.1016/S1387-6473(03)00056-3.

Reis, I., Baron, D., and Shahaf, S. (Jan. 2019). 'Probabilistic Random Forest: A Machine Learning Algorithm for Noisy Data Sets'. In: AJ 157.1, 16, p. 16. DOI: 10.3847/1538-3881/aaf101.

Richter, G. A. (Jan. 1975). 'Search for Optical Identifications in the 5C3 Radio Survey. II. Statistical Treatment and Results'. In: *Astronomische Nachrichten* 296.2, p. 65. DOI: 10.1002/asna.19752960203.

Rohde, D. J., Drinkwater, M. J., Gallagher, M. R., et al. (June 2005). 'Applying machine learning to catalogue matching in astrophysics'. In: MNRAS 360.1, pp. 69–75. DOI: 10.1111/j.1365-2966.2005.08930.x.

Rohde, D. J., Gallagher, M. R., Drinkwater, M. J., and Pimbblet, K. A. (June 2006). 'Matching of catalogues by probabilistic pattern classification'. In: MNRAS 369.1, pp. 2–14. DOI: 10.1111/j.1365-2966.2006.10304.x.

Roscher, R., Bohn, B., Duarte, M. F., and Garcke, J. (2020). 'Explainable Machine Learning for Scientific Insights and Discoveries'. In: *IEEE Access* 8, pp. 42200–42216. DOI: 10.1109/ACCESS.2020.2976199.

Ross, N. P. and Cross, N. J. G. (May 2020). 'The near and mid-infrared photometric properties of known redshift z ≥ 5 quasars'. In: MNRAS 494.1, pp. 789–803. DOI: 10.1093/mnras/staa544.

Ross, N. P., McGreer, I. D., White, M., et al. (Aug. 2013). 'The SDSS-III Baryon Oscillation Spectroscopic Survey: The Quasar Luminosity Function from Data Release Nine'. In: ApJ 773.1, 14, p. 14. DOI: 10.1088/0004-637X/773/1/14.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Statistics. Wiley. ISBN: 9780471087052. URL: https://books.google.com/books?id=0KruAAAAMAAJ.

Rubin, D. B. (Dec. 1976). 'Inference and missing data'. In: *Biometrika* 63.3, pp. 581–592. ISSN: 0006-3444. DOI: 10.1093/biomet/63.3.581.

Saarela, M. and Jauhiainen, S. (Feb. 2021). 'Comparison of feature importance measures as explanations for classification models'. In: *SN Applied Sciences* 3.2, p. 272. ISSN: 2523-3971. DOI: 10.1007/s42452-021-04148-9.

Sabater, J., Best, P. N., Hardcastle, M. J., et al. (Feb. 2019). 'The LoTSS view of radio AGN in the local Universe. The most massive galaxies are always switched on'. In: A&A 622, A17, A17. DOI: 10.1051/0004-6361/201833883.

Sajina, A., Lacy, M., and Pope, A. (June 2022). 'The Past and Future of Mid-Infrared Studies of AGN'. In: *Universe* 8.7, p. 356. DOI: 10.3390/universe8070356.

Salpeter, E. E. (Jan. 1955). 'The Luminosity Function and Stellar Evolution.' In: ApJ 121, p. 161. DOI: 10.1086/145971.

Salvato, M., Buchner, J., Budavári, T., et al. (Feb. 2018). 'Finding counterparts for all-sky X-ray surveys with NWAY: a Bayesian algorithm for cross-matching multiple catalogues'. In: MNRAS 473.4, pp. 4937–4955. DOI: 10.1093/mnras/stx2651.

Salvato, M., Ilbert, O., Hasinger, G., et al. (Dec. 2011). 'Dissecting Photometric Redshift for Active Galactic Nucleus Using XMM- and Chandra-COSMOS Samples'. In: ApJ 742.2, 61, p. 61. DOI: 10.1088/0004-637X/742/2/61.

Salvato, M., Ilbert, O., and Hoyle, B. (June 2019). 'The many flavours of photometric redshifts'. In: *Nature Astronomy* 3, pp. 212–222. DOI: 10.1038/s41550-018-0478-0.

Samuel, A. L. (1959). 'Some Studies in Machine Learning Using the Game of Checkers'. In: *IBM Journal of Research and Development* 3.3, pp. 210–229. DOI: 10.1147/rd.33.0210.

Sánchez-Sáez, P., Reyes, I., Valenzuela, C., et al. (Mar. 2021). 'Alert Classification for the ALeRCE Broker System: The Light Curve Classifier'. In: AJ 161.3, 141, p. 141. DOI: 10.3847/1538-3881/abd5c1.

Santos, M. S., Abreu, P. H., Japkowicz, N., et al. (Dec. 2022). 'On the joint-effect of class imbalance and overlap: a critical review'. In: *Artificial Intelligence Review* 55.8, pp. 6207–6275. ISSN: 1573-7462. DOI: 10.1007/s10462-022-10150-3.

Sartori, L. F., Schawinski, K., Treister, E., et al. (Dec. 2015). 'The search for active black holes in nearby low-mass galaxies using optical and mid-IR data'. In: MNRAS 454.4, pp. 3722–3742. DOI: 10.1093/mnras/stv2238.

Schapire, R. E. (June 1990). 'The strength of weak learnability'. In: *Machine Learning* 5.2, pp. 197–227. ISSN: 1573-0565. DOI: 10.1007/BF00116037.

Schapire, R. E., Freund, Y., Bartlett, P., and Lee, W. S. (1998). 'Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods'. In: *The Annals of Statistics* 26.5, pp. 1651–1686. ISSN: 00905364. URL: http://www.jstor.org/stable/120016 (visited on 05/12/2023).

Schechter, P. (Jan. 1976). 'An analytic expression for the luminosity function for galaxies.' In: ApJ 203, pp. 297–306. DOI: 10.1086/154079.

Schmidt, M. (Feb. 1968). 'Space Distribution and Luminosity Functions of Quasi-Stellar Radio Sources'. In: ApJ 151, p. 393. DOI: 10.1086/149446.

Schneider, P. C., Freund, S., Czesla, S., et al. (May 2022). 'The eROSITA Final Equatorial-Depth Survey (eFEDS). The Stellar Counterparts of eROSITA sources identified by machine learning and Bayesian algorithms'. In: A&A 661, A6, A6. DOI: 10.1051/0004-6361/202141133.

Schuecker, P. (Jan. 1993). 'Automated Galaxy Redshift Measurements from Very Low Dispersion Objective Prism Schmidt Plates'. In: ApJS 84, p. 39. DOI: 10.1086/191744.

Scoville, N., Aussel, H., Brusa, M., et al. (Sept. 2007). 'The Cosmic Evolution Survey (COSMOS): Overview'. In: ApJS 172.1, pp. 1–8. DOI: 10.1086/516585.

Selina, R. J., Murphy, E. J., McKinnon, M., et al. (Dec. 2018). 'The ngVLA Reference Design'. In: *Science with a Next Generation Very Large Array*. Ed. by E. Murphy. Vol. 517. Astronomical Society of the Pacific Conference Series, p. 15. DOI: 10.48550/arXiv.1810.08197.

Selina, R., Murphy, E., and Beasley, A. (Jan. 2023). 'The ngVLA: A Technical Overview'. In: *American Astronomical Society Meeting Abstracts*. Vol. 55. American Astronomical Society Meeting Abstracts, 357.02, p. 357.02.

Sen, S., Agarwal, S., Chakraborty, P., and Singh, K. P. (Feb. 2022). 'Astronomical big data processing using machine learning: A comprehensive review'. In: *Experimental Astronomy* 53.1, pp. 1–43. DOI: 10.1007/s10686-021-09827-4.

Shapley, L. S. (1953). 'A Value for n-Person Games'. In: *Contributions to the Theory of Games (AM-28), Volume II*. Vol. 1. Princeton University Press, pp. 307–318. DOI: 10.1515/9781400881970-018.

Shimwell, T. W., Tasse, C., Hardcastle, M. J., et al. (Feb. 2019). 'The LOFAR Two-metre Sky Survey. II. First data release'. In: A&A 622, A1, A1. DOI: 10.1051/0004-6361/201833559.

Shy, S., Tak, H., Feigelson, E. D., et al. (July 2022). 'Incorporating Measurement Error in Astronomical Object Classification'. In: AJ 164.1, 6, p. 6. DOI: 10.3847/1538-3881/ac6e64.

Silva, L., Schurer, A., Granato, G. L., et al. (Jan. 2011). 'Modelling the spectral energy distribution of galaxies: introducing the artificial neural network'. In: MNRAS 410.3, pp. 2043–2056. DOI: 10.1111/j.1365-2966.2010.17580.x.

Silva Filho, T., Song, H., Perello-Nieto, M., et al. (Sept. 2023). 'Classifier calibration: a survey on how to assess and improve predicted class probabilities'. In: *Machine Learning* 112.9, pp. 3211–3260. ISSN: 1573-0565. DOI: 10.1007/s10994-023-06336-7.

Simpson, C., Rawlings, S., Ivison, R., et al. (Apr. 2012). 'Radio imaging of the Subaru/XMM-Newton Deep Field-III. Evolution of the radio luminosity function beyond z= 1'. In: MNRAS 421.4, pp. 3060–3083. DOI: 10.1111/j.1365-2966.2012.20529.x.

Singh, V., Beelen, A., Wadadekar, Y., et al. (Sept. 2014). 'Multiwavelength characterization of faint ultra steep spectrum radio sources: A search for high-redshift radio galaxies'. In: A&A 569, A52, A52. DOI: 10.1051/0004-6361/201423644.

Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. (Feb. 2006). 'The Two Micron All Sky Survey (2MASS)'. In: AJ 131.2, pp. 1163–1183. DOI: 10.1086/498708.

Šlaus, B., Smolčić, V., Novak, M., et al. (June 2020). 'The XXL Survey. XLI. Radio AGN luminosity functions based on the GMRT 610 MHz continuum observations'. In: A&A 638, A46, A46. DOI: 10.1051/0004-6361/201937258.

Sola, J. and Sevilla, J. (1997). 'Importance of input data normalization for the application of neural networks to complex industrial problems'. In: *IEEE Transactions on Nuclear Science* 44.3, pp. 1464–1468. DOI: 10.1109/23.589532.

Sollich, P. and Krogh, A. (1995). 'Learning with ensembles: How overfitting can be useful'. In: *Advances in Neural Information Processing Systems*. Ed. by D. Touretzky, M. Mozer, and M. Hasselmo. Vol. 8. MIT Press. URL: https://proceedings.neurips.cc/paper_files/paper/1995/file/1019c8091693ef5c5f55970346633f92-Paper.pdf.

Sommer, M. W., Basu, K., Pacaud, F., et al. (May 2011). 'Redshift evolution of the 1.4 GHz volume averaged radio luminosity function in clusters of galaxies'. In: A&A 529, A124, A124. DOI: 10.1051/0004-6361/201016150.

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. URL: https://books.google.pt/books?id=rpS8GAAACAAJ.

Spiniello, C. and Agnello, A. (Oct. 2019). 'VEXAS: VISTA EXtension to Auxiliary Surveys. Data Release 1. The southern Galactic hemisphere'. In: A&A 630, A146, A146. DOI: 10.1051/0004-6361/201936311.

Sravan, N., Graham, M. J., Coughlin, M. W., et al. (July 2023). 'Machine-directed gravitational-wave counterpart discovery'. In: *arXiv e-prints*, arXiv:2307.09213, arXiv:2307.09213. DOI: 10.48550/arXiv.2307.09213.

Steidel, C. C., Adelberger, K. L., Giavalisco, M., et al. (July 1999). 'Lyman-Break Galaxies at z>∼4 and the Evolution of the Ultraviolet Luminosity Density at High Redshift'. In: ApJ 519.1, pp. 1–17. DOI: 10.1086/307363.

Steidel, C. C., Giavalisco, M., Pettini, M., et al. (May 1996). 'Spectroscopic Confirmation of a Population of Normal Star-forming Galaxies at Redshifts Z > 3'. In: ApJ 462, p. L17. DOI: 10.1086/310029.

Steidel, C. C. and Hamilton, D. (Sept. 1992). 'Deep Imaging of redshift QSO Fields Below the Lyman Limit. I. The Field of Q0000-263 and galaxies at Z= 3.4'. In: AJ 104, p. 941. DOI: 10.1086/116287.

Stern, D., Assef, R. J., Benford, D. J., et al. (July 2012). 'Mid-infrared Selection of Active Galactic Nuclei with the Wide-Field Infrared Survey Explorer. I. Characterizing WISE-selected Active Galactic Nuclei in COSMOS'. In: ApJ 753.1, 30, p. 30. DOI: 10.1088/0004-637X/753/1/30.

Stern, D., Eisenhardt, P., Gorjian, V., et al. (Sept. 2005). 'Mid-Infrared Selection of Active Galaxies'. In: ApJ 631.1, pp. 163–168. DOI: 10.1086/432523.

Stone, M. (1974). 'Cross-Validatory Choice and Assessment of Statistical Predictions'. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.2, pp. 111–133. DOI: https://doi.org/10.1111/j.2517-6161.1974.tb00994.x.

Storey-Fisher, K., Hogg, D. W., Rix, H.-W., et al. (June 2023). 'Quaia, the Gaia-unWISE Quasar Catalog: An All-Sky Spectroscopic Quasar Sample'. In: *arXiv e-prints*, arXiv:2306.17749, arXiv:2306.17749. DOI: 10.48550/arXiv.2306.17749.

Storey-Fisher, K., Huertas-Company, M., Ramachandra, N., et al. (Dec. 2021). 'Anomaly detection in Hyper Suprime-Cam galaxy images with generative adversarial networks'. In: MNRAS 508.2, pp. 2946–2963. DOI: 10.1093/mnras/stab2589.

Stoughton, C., Lupton, R. H., Bernardi, M., et al. (Jan. 2002). 'Sloan Digital Sky Survey: Early Data Release'. In: AJ 123.1, pp. 485–548. DOI: 10.1086/324741.

Suganuma, M., Yoshii, Y., Kobayashi, Y., et al. (Mar. 2006). 'Reverberation Measurements of the Inner Radius of the Dust Torus in Nearby Seyfert 1 Galaxies'. In: ApJ 639.1, pp. 46–63. DOI: 10.1086/499326.

Sutherland, W. and Saunders, W. (Dec. 1992). 'On the likelihood ratio for source identification.' In: MNRAS 259, pp. 413–420. DOI: 10.1093/mnras/259.3.413.

Sweijen, F., van Weeren, R. J., Röttgering, H. J. A., et al. (Jan. 2022). 'Deep sub-arcsecond wide-field imaging of the Lockman Hole field at 144 MHz'. In: *Nature Astronomy* 6, pp. 350–356. DOI: 10.1038/s41550-021-01573-z.

Taylor, M. B. (Dec. 2005). 'TOPCAT & STIL: Starlink Table/VOTable Processing Software'. In: *Astronomical Data Analysis Software and Systems XIV*. Ed. by P. Shopbell, M. Britton, and R. Ebert. Vol. 347. Astronomical Society of the Pacific Conference Series, p. 29.

Thomas, N., Davé, R., Jarvis, M. J., and Anglés-Alcázar, D. (May 2021). 'The radio galaxy population in the SIMBA simulations'. In: MNRAS 503.3, pp. 3492–3509. DOI: 10.1093/mnras/stab654.

Thorne, J., Robotham, A., Davies, L., and Bellstedt, S. (Mar. 2022a). *AGN Unification Diagram*. DOI: 10.5281/zenodo.6381013.

Thorne, J. E., Robotham, A. S. G., Davies, L. J. M., et al. (Feb. 2022b). 'Deep Extragalactic VIsible Legacy Survey (DEVILS): identification of AGN through SED fitting and the evolution of the bolometric AGN luminosity function'. In: MNRAS 509.4, pp. 4940–4961. DOI: 10.1093/mnras/stab3208.

Toba, Y., Oyabu, S., Matsuhara, H., et al. (June 2014). 'Luminosity and Redshift Dependence of the Covering Factor of Active Galactic Nuclei viewed with WISE and Sloan Digital Sky Survey'. In: ApJ 788.1, 45, p. 45. DOI: 10.1088/0004-637X/788/1/45.

Tonry, J. and Davis, M. (Oct. 1979). 'A survey of galaxy redshifts. I. Data reduction techniques.' In: AJ 84, pp. 1511–1525. DOI: 10.1086/112569.

Toth, M. J., Goran, M. I., Ades, P. A., et al. (1993). 'Examination of data normalization procedures for expressing peak VO2 data'. In: *Journal of applied physiology* 75.5, pp. 2288–2292.

Tripodi, R., Feruglio, C., Fiore, F., et al. (Sept. 2022). 'Black hole and host galaxy growth in an isolated z ∼ 6 QSO observed with ALMA'. In: A&A 665, A107, A107. DOI: 10.1051/0004-6361/202243920.

Troyer, J., Starkey, D., Cackett, E. M., et al. (Mar. 2016). 'Correlated X-ray/ultraviolet/optical variability in NGC 6814'. In: MNRAS 456.4, pp. 4040–4050. DOI: 10.1093/mnras/stv2862.

Urry, C. M. and Padovani, P. (Sept. 1995). 'Unified Schemes for Radio-Loud Active Galactic Nuclei'. In: PASP 107, p. 803. DOI: 10.1086/133630.

Uttley, P., Edelson, R., McHardy, I. M., et al. (Feb. 2003). 'Correlated Long-Term Optical and X-Ray Variations in NGC 5548'. In: ApJ 584.2, pp. L53–L56. DOI: 10.1086/373887.

Uzgil, B. D., Oesch, P. A., Walter, F., et al. (May 2021). 'The ALMA Spectroscopic Survey in the HUDF: A Search for [C II] Emitters at 6 ≤ z ≤ 8'. In: ApJ 912.1, 67, p. 67. DOI: 10.3847/1538-4357/abe86b.

Van Calster, B., McLernon, D. J., Smeden, M. van, et al. (Dec. 2019). 'Calibration: the Achilles heel of predictive analytics'. In: *BMC Medicine* 17.1, p. 230. ISSN: 1741-7015. DOI: 10.1186/s12916-019-1466-7.

van der Velden, E. (Feb. 2020). 'CMasher: Scientific colormaps for making accessible, informative and 'cmashing' plots'. In: *The Journal of Open Source Software* 5.46, 2004, p. 2004. DOI: 10.21105/joss.02004.

van der Vlugt, D., Hodge, J. A., Algera, H. S. B., et al. (Dec. 2022). 'An Ultra-deep Multiband Very Large Array (VLA) Survey of the Faint Radio Sky (COSMOS-XS): New Constraints on the Cosmic Star Formation History'. In: ApJ 941.1, 10, p. 10. DOI: 10.3847/1538-4357/ac99db.

van Haarlem, M. P., Wise, M. W., Gunst, A. W., et al. (July 2013). 'LOFAR: The LOw-Frequency ARray'. In: A&A 556, A2, A2. DOI: 10.1051/0004-6361/201220873.

van Rijsbergen, C. J. (1979). *Information Retrieval*. 2nd. USA: Butterworth-Heinemann. ISBN: 0408709294.

Vanden Berk, D. E., Wilhite, B. C., Kron, R. G., et al. (Feb. 2004). 'The Ensemble Photometric Variability of ~25,000 Quasars in the Sloan Digital Sky Survey'. In: ApJ 601.2, pp. 692–714. DOI: 10.1086/380563.

Vanschoren, J. (2019). 'Meta-Learning'. In: *Automated Machine Learning: Methods, Systems, Challenges*. Ed. by F. Hutter, L. Kotthoff, and J. Vanschoren. Cham: Springer International Publishing, pp. 35–61. ISBN: 978-3-030-05318-5. DOI: 10.1007/978-3-030-05318-5_2.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer. ISBN: 9780387945590. URL: https://books.google.pt/books?id=r_ayQgAACAAJ.

Vardoulaki, E., Jiménez Andrade, E. F., Delvecchio, I., et al. (Apr. 2021). 'FR-type radio sources at 3 GHz VLA-COSMOS: Relation to physical properties and large-scale environment'. In: A&A 648, A102, A102. DOI: 10.1051/0004-6361/202039488.

Villaescusa-Navarro, F., Anglés-Alcázar, D., Genel, S., et al. (July 2021). 'The CAMELS Project: Cosmology and Astrophysics with Machine-learning Simulations'. In: ApJ 915.1, 71, p. 71. DOI: 10.3847/1538-4357/abf7ba.

Vuttipittayamongkol, P., Elyan, E., and Petrovski, A. (2021). 'On the class overlap problem in imbalanced data classification'. In: *Knowledge-Based Systems* 212, p. 106631. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2020.106631.

Wagstaff, K. L., Huff, E., and Rebbapragada, U. (July 2022). 'Machine-Assisted Discovery Through Identification and Explanation of Anomalies in Astronomical Surveys'. In: *Astronomical Society of the Pacific Conference Series*. Ed. by J. E. Ruiz, F. Pierfedereci, and P. Teuben. Vol. 532. Astronomical Society of the Pacific Conference Series, p. 183.

Walcher, J., Groves, B., Budavári, T., and Dale, D. (Jan. 2011). 'Fitting the integrated spectral energy distributions of galaxies'. In: Ap&SS 331, pp. 1–52. DOI: 10.1007/s10509-010-0458-z.

Wenzl, L., Schindler, J.-T., Fan, X., et al. (Aug. 2021). 'Random Forests as a Viable Method to Select and Discover High-redshift Quasars'. In: AJ 162.2, 72, p. 72. DOI: 10.3847/1538-3881/ac0254.

Werner, M. W., Roellig, T. L., Low, F. J., et al. (Sept. 2004). 'The Spitzer Space Telescope Mission'. In: ApJS 154.1, pp. 1–9. DOI: 10.1086/422992.

Whittam, I. H., Prescott, M., Hale, C. L., et al. (Jan. 2024). 'MIGHTEE: Multi-wavelength counterparts in the COSMOS field'. In: MNRAS 527.2, pp. 3231–3245. DOI: 10.1093/mnras/stad3307.

Wilber, A. G., Dabbech, A., Terris, M., et al. (July 2023). 'Scalable precision wide-field imaging in radio interferometry - II. AIRI validated on ASKAP data'. In: MNRAS 522.4, pp. 5576–5587. DOI: 10.1093/mnras/stad1353.

Williams, W. L., Calistro Rivera, G., Best, P. N., et al. (Apr. 2018). 'LOFAR-Boötes: properties of high- and low-excitation radio galaxies at $0.5 < z < 2.0$'. In: MNRAS 475.3, pp. 3429–3452. DOI: 10.1093/mnras/sty026.

Williams, W. L. and Röttgering, H. J. A. (June 2015). 'Radio-AGN feedback: when the little ones were monsters'. In: MNRAS 450.2, pp. 1538–1545. DOI: 10.1093/mnras/stv692.

Wolpert, D. H. (1992). 'Stacked generalization'. In: *Neural Networks* 5.2, pp. 241–259. ISSN: 0893-6080. DOI: https://doi.org/10.1016/S0893-6080(05)80023-1.

Wolstencroft, R. D., Savage, A., Clowes, R. G., et al. (Nov. 1986). 'The identification of IRAS point sources- I. A 304 deg2 field centred on the South Galactic Pole.' In: MNRAS 223, pp. 279–302. DOI: 10.1093/mnras/223.2.279.

Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. (Dec. 2010). 'The Wide-field Infrared Survey Explorer (WISE): Mission Description and Initial On-orbit Performance'. In: AJ 140.6, pp. 1868–1881. DOI: 10.1088/0004-6256/140/6/1868.

Wu, C., Wong, O. I., Rudnick, L., et al. (Jan. 2019). 'Radio Galaxy Zoo: CLARAN - a deep learning classifier for radio morphologies'. In: MNRAS 482.1, pp. 1211–1230. DOI: 10.1093/mnras/sty2646.

Xue, Y. Q., Luo, B., Brandt, W. N., et al. (July 2011). 'The Chandra Deep Field-South Survey: 4 Ms Source Catalogs'. In: ApJS 195.1, 10, p. 10. DOI: 10.1088/0067-0049/195/1/10.

Yan, W., Brandt, W. N., Zou, F., et al. (July 2023). 'The Most Obscured AGNs in the XMM-SERVS Fields'. In: ApJ 951.1, 27, p. 27. DOI: 10.3847/1538-4357/accea6.

Yang, J. (Sept. 2021). 'Fast TreeSHAP: Accelerating SHAP Value Computation for Trees'. In: *arXiv e-prints*, arXiv:2109.09847, arXiv:2109.09847. DOI: 10.48550/arXiv.2109.09847.

Yang, Q. and Shen, Y. (Jan. 2023). 'A Southern Photometric Quasar Catalog from the Dark Energy Survey Data Release 2'. In: ApJS 264.1, 9, p. 9. DOI: 10.3847/1538-4365/ac9ea8.

Ye, H., Sweijen, F., van Weeren, R., et al. (Sept. 2023). '1-arcsecond imaging strategy for the LoTSS survey using the International LOFAR Telescope'. In: *arXiv e-prints*, arXiv:2309.16560, arXiv:2309.16560. DOI: 10.48550/arXiv.2309.16560.

Yeo, I.-K. and Johnson, R. A. (Dec. 2000). 'A new family of power transformations to improve normality or symmetry'. In: *Biometrika* 87.4, pp. 954–959. ISSN: 0006-3444. DOI: 10.1093/biomet/87.4.954.

Yerushalmy, J. (1947). 'Statistical Problems in Assessing Methods of Medical Diagnosis, with Special Reference to X-Ray Techniques'. In: *Public Health Reports (1896-1970)* 62.40, pp. 1432–1449. ISSN: 00946214. URL: http://www.jstor.org/stable/4586294 (visited on 10/08/2022).

York, D. G., Adelman, J., Anderson John E., J., et al. (Sept. 2000). 'The Sloan Digital Sky Survey: Technical Summary'. In: AJ 120.3, pp. 1579–1587. DOI: 10.1086/301513.

Yuan, Z., Jarvis, M. J., and Wang, J. (May 2020). 'A Flexible Method for Estimating Luminosity Functions via Kernel Density Estimation'. In: ApJS 248.1, 1, p. 1. DOI: 10.3847/1538-4365/ab855b.

Yuan, Z., Zhang, X., Wang, J., et al. (May 2022). 'A Flexible Method for Estimating Luminosity Functions via Kernel Density Estimation. II. Generalization and Python Implementation'. In: ApJS 260.1, 10, p. 10. DOI: 10.3847/1538-4365/ac596a.

Yule, G. U. (1912). 'On the Methods of Measuring Association Between Two Attributes'. In: *Journal of the Royal Statistical Society* 75.6, pp. 579–652. ISSN: 09528385. URL: http://www.jstor.org/stable/2340126.

Zajaček, M., Busch, G., Valencia-S., M., et al. (Oct. 2019). 'Radio spectral index distribution of SDSS-FIRST sources across optical diagnostic diagrams'. In: A&A 630, A83, A83. DOI: 10.1051/0004-6361/201833388.

Zammit, M. A. and Adami, K. Z. (Nov. 2023). 'Machine Learning Applications in Jupiter-host Star Classification using Stellar Spectra'. In: MNRAS. DOI: 10.1093/mnras/stad3668.

Zhang, Y. and Zhao, Y. (May 2015). 'Astronomy in the Big Data Era'. In: *Data Science Journal* 14, p. 11. DOI: 10.5334/dsj-2015-011.

Zheng, A. and Casari, A. (2018). *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly. ISBN: 9781491953242. URL: https://books.google.pt/books?id=Ho0UvgAACAAJ.

Zitlau, R., Hoyle, B., Paech, K., et al. (Aug. 2016). 'Stacking for machine learning redshifts applied to SDSS galaxies'. In: MNRAS 460.3, pp. 3152–3162. DOI: 10.1093/mnras/stw1454.

Zou, H., Gao, J., Zhou, X., and Kong, X. (May 2019). 'Photometric Redshifts and Stellar Masses for Galaxies from the DESI Legacy Imaging Surveys'. In: ApJS 242.1, 8, p. 8. DOI: 10.3847/1538-4365/ab1847.

This page intentionally left blank.

Appendices

This page intentionally left blank.

A

Sample of predicted radio-detectable AGN

The columns shown in the prediction results for sources in both HETDEX and S82 fields are described in Table A.1. Full datasets and models from prediction pipeline can be obtained from Carvajal et al. (2023b) at https://zenodo.org/doi/10.5281/zenodo.10220008.

Table A.1: Table columns descriptions.

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 source as it appears in CW catalogue
band_num	Number of non-radio bands with valid measurement per source (cf. Sect. 5.5)
class	1 if source is a confirmed AGN by MQC. 0 if has been spectroscopically confirmed as galaxy in SDSS-DR16. Sources with no value do not have a spectroscopic classification in this catalogue
Sint_LOFAR (or Fint_VLAS82)[a]	Imputed integrated flux (in mJy) of source from LOFAR or VLAS82
Sint_LOFAR_non_imp (or Fint_VLAS82_non_imp)[a]	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 meta AGN-galaxy classifier for prediction to be an AGN
Prob_AGN	Probability from calibrated meta AGN-galaxy classifier for prediction to be AGN
LOFAR_detect	1 if source has been detected on LoTSS survey or in their analogue surveys for fields different to HETDEX (see Sects. 5.1 and 5.2). 0 otherwise
Score_radio_AGN	Score from meta radio detection model for prediction to be detected in radio
Prob_radio_AGN	Probability from calibrated radio detection model for prediction to be detected in radio
radio_AGN	class × LOFAR_detect. 1 if source is AGN and has been detected in radio. 0 otherwise
Score_rAGN	Score_AGN × Score_radio. Score of source for it to be AGN detected in radio
Prob_rAGN	Prob_AGN × Prob_radio. Probability of source for it to be AGN detected in radio
Z	Spectroscopic redshift as listed by the MQC (if available)
pred_Z	Redshift value predicted by our model

[a] Sources from HETDEX field have columns with LOFAR suffix. Sources from VLAS82 have VLAS82 suffix.

Table A.2: Predicted and original properties for 20 sources in testing subset with highest redshift predicted radio AGN. Sources sorted by decreasing predicted redshift.

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
9 898 717	203.016113	55.518097	9	1.0	0.500082	0.954114	0	0.390861	0.375122	0.195462	0.357909	4.738	4.3679
168 686	164.769135	45.806320	8	1.0	0.500048	0.858157	0	0.450279	0.418719	0.225161	0.359326	4.893	4.1733
14 437 074	213.226517	54.236343	9	1.0	0.500090	0.965187	0	0.251632	0.263746	0.125839	0.254564	4.326	4.0475
10 408 176	188.163651	52.880898	9	1.0	0.500012	0.622448	0	0.604838	0.526003	0.302426	0.327410	4.340	3.9553
12 612 753	227.216370	51.941029	9	1.0	0.500055	0.887909	0	0.364423	0.355080	0.182231	0.315278	3.795	3.8797
5 283 988	210.532974	49.500351	9	1.0	0.500095	0.971223	0	0.188993	0.207802	0.094514	0.201822	3.800	3.8733
6 005 721	188.469818	47.834412	9	1.0	0.500045	0.845762	0	0.221281	0.237188	0.110651	0.200604	3.795	3.7482
13 284 852	218.364182	48.398373	9	1.0	0.500086	0.961052	0	0.340675	0.336683	0.170367	0.323570	3.892	3.7222
12 308 398	222.563293	49.665272	8	1.0	0.500061	0.906658	1	0.385996	0.371466	0.193021	0.336793	2.059	3.7173
8 147 562	200.292435	50.367386	9	1.0	0.500090	0.965431	1	0.406555	0.386820	0.203314	0.373448	3.711	3.6748
4 433 144	215.540436	46.992298	9	1.0	0.500085	0.958811	1	0.363854	0.354644	0.181958	0.340036	3.812	3.6541
12 285 636	221.428192	49.046921	12	1.0	0.500037	0.803828	1	0.254536	0.266237	0.127277	0.214009	3.880	3.5737
8 170 701	200.040466	50.849586	9	1.0	0.500040	0.818872	0	0.232864	0.247439	0.116441	0.202620	3.827	3.5436
12 118 249	223.858124	48.950409	9	1.0	0.500049	0.862100	1	0.280344	0.288020	0.140186	0.248302	3.501	3.5381
1 537 357	177.051697	52.934574	9	1.0	0.500015	0.645263	0	0.240127	0.253793	0.120067	0.163763	3.885	3.5037
10 731 551	190.228714	54.614464	9	1.0	0.500055	0.887182	1	0.277223	0.285418	0.138627	0.253218	3.938	3.4858
8 721 103	203.096924	50.575378	9	1.0	0.500093	0.969007	0	0.267817	0.277525	0.133933	0.268924	3.832	3.4604
10 679 277	190.777359	53.089466	9	1.0	0.500104	0.978721	1	0.323977	0.323506	0.162022	0.316622	3.575	3.3764
8 434 867	194.881256	52.609665	9	1.0	0.500066	0.920703	0	0.811700	0.672552	0.405903	0.619221	0.374	3.1584
10 314 475	188.218765	51.662666	9	1.0	0.500109	0.982703	0	0.545364	0.485399	0.272741	0.477003	2.157	3.0701

Table A.3: Predicted and original properties for the 20 sources in S82 with highest predicted redshift on labelled sources predicted to be radio AGN. Sources sorted by decreasing predicted redshift.

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
1 406 323	32.679794	-0.305035	6	1.0	0.500050	0.866373	1	0.185842	0.204867	0.092930	0.177491	4.650	4.4986
326 139	33.580879	-1.121398	8	1.0	0.500040	0.822622	0	0.208769	0.225946	0.104393	0.185868	4.600	4.3785
633 752	12.526446	-0.888660	9	1.0	0.500035	0.793882	0	0.206182	0.223600	0.103098	0.177512	4.310	4.2946
2 834 844	344.101440	0.789000	7	1.0	0.500062	0.909395	0	0.375735	0.363709	0.187891	0.330756	4.099	4.0635
3 191 865	31.881712	1.063655	9	1.0	0.500087	0.962260	0	0.264210	0.274477	0.132128	0.264118	3.841	4.0509
834 857	21.023815	-0.735512	9	1.0	0.500080	0.951495	0	0.270463	0.279754	0.135253	0.266185	4.073	4.0476
927 403	30.860348	-0.664943	9	1.0	0.500086	0.960503	0	0.237843	0.251800	0.118942	0.241855	4.158	3.9690
1 760 668	343.580109	-0.031951	9	1.0	0.500088	0.962655	0	0.262341	0.272891	0.131193	0.262700	3.710	3.9594
1 539 341	35.058613	-0.202904	5	1.0	0.500028	0.746304	0	0.243109	0.256385	0.121561	0.191341	3.712	3.9175
794 288	21.752874	-0.766390	12	1.0	0.500044	0.838993	0	0.830699	0.687772	0.415386	0.577036	4.105	3.8631
201 543	18.192499	-1.218799	9	1.0	0.500062	0.909994	0	0.383420	0.369525	0.191734	0.336265	3.589	3.7324
2 705 463	27.703466	0.690627	12	1.0	0.500092	0.968344	0	0.284506	0.291475	0.142279	0.282248	3.685	3.6814
567 008	13.804744	-0.939195	9	1.0	0.500114	0.985173	0	0.473856	0.435550	0.236982	0.429092	3.622	3.5531
2 610 655	334.937775	0.618921	9	1.0	0.500107	0.981287	0	0.201475	0.219309	0.100759	0.215205	3.531	3.5031
2 972 720	341.585175	0.894723	9	1.0	0.500087	0.961458	0	0.752861	0.628400	0.376496	0.604180	3.676	3.4776
315 550	342.345581	-1.129457	9	1.0	0.500070	0.930942	0	0.485239	0.443596	0.242653	0.412962	3.989	3.4724
2 774 066	21.015764	0.742418	12	1.0	0.500045	0.843852	0	0.642026	0.551288	0.321042	0.465205	3.836	3.4492
3 301 527	346.665192	1.148655	9	1.0	0.500045	0.846709	0	0.305069	0.308323	0.152548	0.261060	3.648	3.4407
1 926 483	22.206123	0.095463	9	1.0	0.500117	0.986782	0	0.427273	0.402055	0.213687	0.396741	3.746	3.4282
211 079	31.115847	-1.211047	9	1.0	0.500106	0.980813	0	0.455339	0.422351	0.227718	0.414247	3.940	3.3345

Table A.4: Predicted and original properties for the 20 sources in HETDEX field with highest predicted redshift on the unlabelled sources predicted to be radio AGN.

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
9 544 254	201.309235	53.746429	6	0.500007	0.578804	0	0.351672	0.345250	0.175838	0.199832	4.7114
12 355 845	220.838120	50.319016	5	0.500007	0.578804	0	0.937123	0.794128	0.468568	0.459644	4.6064
13 814 216	219.839142	52.660328	7	0.500015	0.650248	0	0.213846	0.230529	0.106926	0.149901	4.5622
6 698 239	184.694901	49.063766	5	0.499995	0.467527	0	0.799085	0.662753	0.399538	0.309855	4.5483
2 951 011	175.882446	55.497799	5	0.500008	0.589419	0	0.823295	0.681768	0.411654	0.401847	4.5320
7 281 571	202.589035	47.134212	9	0.500032	0.773522	0	0.567045	0.500257	0.283541	0.386960	4.5220
8 634 951	206.688446	51.339924	7	0.499981	0.349392	0	0.289692	0.295758	0.144841	0.103336	4.5111
5 075 789	207.003540	48.740067	7	0.499998	0.492974	0	0.400019	0.381966	0.200009	0.188299	4.5087
12 750 738	222.945663	52.196320	7	0.500009	0.593819	0	0.238322	0.252219	0.119163	0.149772	4.5064
1 775 496	163.331070	47.809532	7	0.500032	0.772884	0	0.764361	0.636758	0.382205	0.492140	4.4930
14 577 806	213.842804	56.349747	8	0.500047	0.856822	0	0.615253	0.533082	0.307656	0.456756	4.4663
9 817 016	198.035965	55.984398	5	0.500010	0.603443	0	0.466084	0.430028	0.233047	0.259497	4.4438
12 429 128	226.345795	49.744789	7	0.500007	0.581465	0	0.485468	0.443757	0.242737	0.258029	4.4438
14 500 121	211.942429	55.213791	7	0.500008	0.587655	0	0.809995	0.671214	0.405004	0.394443	4.4355
2 166 668	163.079010	53.084713	7	0.500010	0.605185	0	0.743381	0.621584	0.371698	0.376173	4.4353
7 077 234	181.246124	53.378967	7	0.500012	0.623304	0	0.264222	0.274486	0.132114	0.171088	4.4209
14 074 782	218.186417	54.400047	7	0.500019	0.681823	0	0.396363	0.379240	0.198189	0.258574	4.4065
5 789 667	187.343811	47.441658	8	0.500023	0.706552	0	0.725299	0.608743	0.362666	0.430109	4.4027
11 070 539	183.812180	54.915974	7	0.500012	0.621592	0	0.202662	0.220393	0.101333	0.136995	4.3959
8 943 186	202.578445	53.892410	7	0.499983	0.362751	0	0.331269	0.329286	0.165629	0.119449	4.3951

Table A.5: Predicted and original properties for the 20 sources in S82 with highest predicted redshift on the unlabelled sources predicted to be radio AGN.

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
3 244 450	26.276423	1.104065	7	0.500002	0.531172	0	0.542061	0.483128	0.271031	0.256624	4.3938
1 062 270	11.744675	-0.562642	7	0.499982	0.356043	0	0.196326	0.214586	0.098159	0.076402	4.3563
3 261 269	28.882526	1.117103	7	0.500011	0.608660	0	0.354936	0.347777	0.177472	0.211678	4.3153
1 466 227	18.157259	-0.258997	5	0.500013	0.630968	0	0.456207	0.422973	0.228110	0.266882	4.3146
1 134 866	11.304936	-0.507943	7	0.500011	0.616439	0	0.226178	0.241539	0.113091	0.148894	4.3140
1 726 140	29.144888	-0.058600	7	0.499984	0.369512	0	0.261730	0.272373	0.130861	0.100645	4.3125
220 011	331.973877	-1.203849	7	0.500016	0.656024	0	0.734737	0.615421	0.367380	0.403731	4.3098
2 005 587	30.130974	0.155751	7	0.500017	0.665006	0	0.828544	0.686014	0.414286	0.456203	4.2972
1 058 229	344.614166	-0.565701	5	0.500015	0.645263	0	0.394774	0.378052	0.197393	0.243943	4.2899
2 914 632	344.342316	0.850049	7	0.500009	0.594697	0	0.601936	0.524030	0.300973	0.311639	4.2767
1 255 517	20.319715	-0.418364	7	0.500012	0.625012	0	0.210549	0.227556	0.105277	0.142226	4.2711
762 162	340.426544	-0.790798	7	0.500014	0.638566	0	0.671060	0.571113	0.335539	0.364693	4.2687
1 167 955	21.433058	-0.483463	8	0.500016	0.659303	0	0.776046	0.645368	0.388036	0.425493	4.2652
1 160 900	35.526070	-0.488795	7	0.500013	0.630119	0	0.614667	0.532684	0.307341	0.335654	4.2516
706 252	332.367462	-0.833645	5	0.500013	0.630119	0	0.517853	0.466400	0.258933	0.293888	4.2481
510 569	10.896913	-0.981369	7	0.500016	0.651903	0	0.192197	0.210775	0.096102	0.137405	4.2472
2 512 910	333.612640	0.544506	5	0.500014	0.636883	0	0.388904	0.373653	0.194457	0.237974	4.2471
840 809	341.886383	-0.730991	7	0.500042	0.827874	0	0.212097	0.228954	0.106057	0.189545	4.2433
2 932 746	35.531948	0.863913	8	0.500013	0.631815	0	0.917725	0.770169	0.458874	0.486604	4.2421
325 442	344.384644	-1.121920	7	0.500035	0.793285	0	0.624086	0.539086	0.312065	0.427649	4.2353

B

Extended prediction pipeline

For the calculations of Chapter 9, a new instance of the prediction pipeline was trained and implemented. Its steps were kept similar to its original definition (Chapter 4) with some relevant differences. The first difference is in the overall structure of the pipeline. The prediction for radio-detected AGN works as presented in Fig. 4.1, but a new branch has been added for the treatment of radio-detected galaxies (i.e. extragalactic sources without indications of AGN emission). The new branch replicates the steps of the original process. Its stages are presented, graphically, in the flowchart of Fig. B.1.

The difference in the processing of the candidates start when sources are predicted to be galaxies (i.e. not AGN). Instead of being discarded, they are subject to a series of models that replicate the process for predicted AGN. Thus, a new step that predicts their likelihood of being radio detectable is applied. The predicted radio-detectable galaxies have, then, their photometric redshifts predicted. Finally, the results from both branches, AGN and galaxies, are compiled into one single catalogue.

The internal processing of each model remains the same as described in Chapter 5 and Fig. 5.6. The only difference can be found in the data collection. In particular the cross-match of the CW-detected sources with the radio detections (for the training stages, from HETDEX). Instead of using a search radius of 1.″1 (as it is maintained for all the remaining ancillary catalogues), sources are cross-matched with a radius of 6″. This increase of more than five times in distance (and close to 30 times in area) can be explained by the need of obtaining a larger fraction of sources with a radio counterpart (as the results of the cross-match itself would be assessed, Sect. 9.3). A modified version of Table 5.3, with the new search radius for radio counterparts, is shown in Table B.1.

As expected, the number of non-radio counterparts remains the same. In contrast, the number of radio cross-matches has grown by more than a 100 % from the use of a 1.″1 search radius. This change can, in turn, change the metrics for the radio detection models and those of the redshift predictions as well given the modification of the distribution of values of the feature

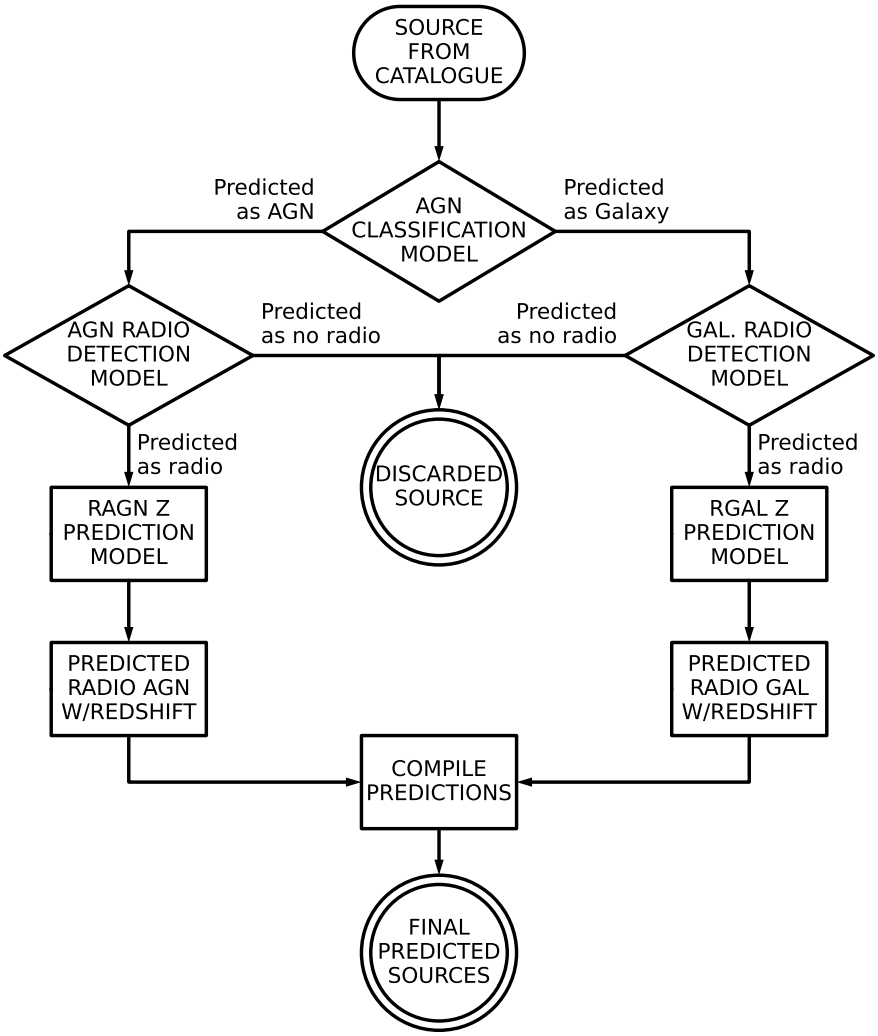


Figure B.1: Flowchart representing the proposed extended prediction pipeline used to predict the presence of radio-detected AGN and galaxies and their redshift values from IR-detected sources.

`radio_detect`. For our purposes, the differences in the metrics for this new set of models do not imply a degradation of the results they provide.

B.1 AGN-galaxy classifier

Results and metrics of the AGN-galaxy classification model have not changed with the extension of the prediction pipeline and the inclusion of a larger number of radio counterparts. This prediction stage does not use any form of radio information for its training and predictions. We refer, then, the reader to Chapter 6 and Sect. 7.1 for the analysis of its results.

Table B.1: Composition of initial catalogue and number of cross matches with additional surveys and catalogues for training of updated pipeline

Survey	HETDEX	Stripe82
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 (6″)	382 431	...
VLAS82 (6″)	...	17 706
MQC (AGN)	50 538	17 743
SDSS (Galaxy)	68 196	4085

Table B.2: Best performing modified models the radio detection classification for AGN

Model	F_β (×100)	MCC (×100)	Precision (×100)	Recall (×100)	Rank
`XGBoost`	36.41 ± 1.82	34.47 ± 2.16	61.93 ± 3.18	27.18 ± 1.53	3.25
`GBC`	36.98 ± 2.20	36.39 ± 2.26	66.47 ± 3.39	27.10 ± 1.99	1.25
`RF`	36.74 ± 1.99	35.90 ± 1.90	65.48 ± 1.91	26.98 ± 1.78	2.25
`CatBoost`	36.18 ± 1.64	35.21 ± 1.68	64.60 ± 2.46	26.55 ± 1.48	3.50
`ET`	35.08 ± 1.09	34.32 ± 1.48	64.22 ± 2.48	25.52 ± 0.83	4.75
`No-skill`	15.29 ± 0.76	0.00 ± 0.91	15.29 ± 0.76	15.29 ± 0.76	6.00

[a] Values and uncertainties as in Table 6.1.

B.2 Radio detection classifiers

The model for the classification of radio-detectable AGN is modified by the change in the distribution of values in its target feature, `radio_detect`. Nevertheless, the features selected for training remain the same as with the original model (i.e. `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`). For the selection of the meta and base models, `XGBoost` was adopted as the meta learner while `GBC`, `RF`, `CatBoost`, and `ET` have been used as base learners. The results of such selection are displayed in Table B.2.

The training of the model for the prediction of radio detectability in predicted galaxies led to the use of 18 features (`W4mag`, `Kmag`, `g_r`, `g_W2`, `r_i`, `r_y`, `i_z`, `i_y`, `z_y`, `z_W2`, `y_J`, `y_W2`, `J_H`, `H_K`, `H_W3`, `W1_W2`, `W1_W3`, and `W3_W4`) together with its target, `radio_detect`. Additionally, and as seen in Table B.3, `RF` was selected as meta learner and `CatBoost`, `XGBoost`, `ET`, and `GBC`, as base models.

Table B.3: Best performing modified models the radio detection classification for galaxies

Model	F_β (×100)	MCC (×100)	Precision (×100)	Recall (×100)	Rank
RF	39.39 ± 2.49	41.69 ± 2.36	76.09 ± 2.42	28.19 ± 2.12	1.25
CatBoost	39.24 ± 2.36	41.53 ± 2.39	75.90 ± 2.69	28.06 ± 1.97	2.25
XGBoost	38.64 ± 1.98	40.38 ± 2.23	73.55 ± 2.87	27.76 ± 1.57	4.00
ET	38.35 ± 2.57	40.69 ± 2.50	75.28 ± 2.78	27.31 ± 2.17	4.00
GBC	37.69 ± 2.32	40.83 ± 2.40	77.39 ± 3.01	26.49 ± 1.89	3.50
No-skill	14.22 ± 1.13	0.00 ± 1.32	14.22 ± 1.14	14.22 ± 1.12	6.00

[a] Values and uncertainties as in Table 6.1.

Table B.4: Results of initial fit for redshift value prediction on predicted radio detectable AGN

Model	σ_{MAD} (×100)	σ_{NMAD} (×100)	σ_z (×100)	σ_z^N (×100)	η (×100)	Rank
RF	17.54 ± 1.24	7.70 ± 0.34	41.55 ± 4.38	19.63 ± 2.57	18.30 ± 1.60	2.0
ET	18.79 ± 1.22	8.32 ± 0.39	40.81 ± 3.25	18.29 ± 1.86	19.56 ± 1.78	2.0
CatBoost	21.07 ± 1.15	9.97 ± 0.25	39.65 ± 2.52	18.06 ± 1.57	21.14 ± 2.17	2.2
XGBoost	22.92 ± 1.20	10.49 ± 0.54	42.16 ± 3.85	19.36 ± 2.16	23.53 ± 1.62	3.8
GBR	28.33 ± 1.48	13.24 ± 0.73	44.77 ± 3.59	20.21 ± 1.93	29.90 ± 1.82	5.0
No-skill	97.06 ± 4.52	39.95 ± 1.89	86.78 ± 1.92	48.17 ± 1.10	72.49 ± 1.96	6.0

[a] Algorithms sorted by increasing σ_{MAD} values.
[b] Uncertainties as in Table 6.1.

B.3 Redshift predictors

The model for the prediction of photometric redshifts of radio-detectable AGN is also modified but, in this case, by the change in number of sources in the training set, which is increased (by the larger search radius for radio-detected sources). Thus, the features selected for training are 17 (i.e. band_num, W4mag, g_r, g_W1, 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 W1_W3). Despite three algorithms having the same mean rank, the algorithm selected to be meta learner is RF given that it presents the best value of σ_{NMAD}, which is the metric to be optimised during training. Its metrics (together with those from the base models) are presented in Table B.4.

The model for the prediction of photometric redshifts of radio-detectable galaxies is also modified and selects 14 features for its training: W4mag, g_r, r_i, r_z, i_z, i_y, z_y, y_J, y_W2, J_H, H_K, K_W3, W1_W2, and W1_W3. The selected algorithm to be meta learner is ET (again, because it presents the best value of σ_{NMAD}) and its metrics (together with those from the base models and an no-skill prediction) are presented in Table B.5.

Table B.5: Results of initial fit for redshift value prediction on predicted radio detectable galaxies

Model	σ_{MAD} (×100)	σ_{NMAD} (×100)	σ_z (×100)	σ_z^{N} (×100)	η (×100)	Rank
ET	3.85 ± 0.25	2.81 ± 0.18	9.89 ± 0.81	7.03 ± 1.06	2.91 ± 0.77	2.0
RF	3.89 ± 0.12	2.86 ± 0.12	9.78 ± 0.88	6.98 ± 1.10	3.08 ± 0.88	2.0
CatBoost	4.01 ± 0.20	2.96 ± 0.12	9.75 ± 0.72	6.99 ± 1.01	2.78 ± 0.56	2.0
XGBoost	4.31 ± 0.27	3.16 ± 0.17	9.98 ± 0.90	7.05 ± 1.05	3.13 ± 0.73	4.2
GBR	4.81 ± 0.16	3.54 ± 0.08	9.96 ± 0.76	7.05 ± 0.99	3.45 ± 1.05	4.6
No-skill	33.51 ± 1.86	21.70 ± 1.31	26.74 ± 0.67	20.94 ± 0.42	49.08 ± 1.93	6.0

[a] Algorithms sorted by increasing σ_{MAD} values.
[b] Uncertainties as in Table 6.1.

B.4 Pipeline prediction

Table B.6: Hyper-parameters values for meta-learners in modified pipeline 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 for AGN (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		
Radio detection model for galaxies (RF)			
Parameter	Value	Parameter	Value
n_estimators	17	max_depth	6
min_impurity_decrease	0.0000	max_features	0.4280
bootstrap	False	criterion	gini
class_weight	balanced_subsample	min_samples_split	10
min_samples_leaf	3		
Redshift prediction model for rAGN (RF)			
Parameter	Value	Parameter	Value
n_estimators	187	max_depth	9
min_impurity_decrease	0.0000	max_features	0.6346
bootstrap	False	criterion	mae
min_samples_split	3	min_samples_leaf	5
Redshift prediction model for rGal (ET)			
Parameter	Value	Parameter	Value
n_estimators	100	criterion	mse
max_depth	None	min_impurity_decrease	0.0000
max_features	auto	bootstrap	False
min_samples_split	2	min_samples_leaf	1

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