UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE FÍSICA

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

MMXXIII

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE FÍSICA

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

MMXXIII

This page intentionally left blank.

ACKNOWLEDGEMENTS

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

I acknowledge 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).

This thesis would have not been possible without the help of that person who bought all the food we needed.

Part of this work was funded by the Big World Fund, which, in its second call, provided us with all the needed materials.

This work uses data from our local library. We thank their kind help with the data we needed.

This research made use of many different computer codes which we will not enumerate for space reasons.

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

This page intentionally left blank.

RESUMO

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.

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

This page intentionally left blank.

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.

RESUMO ALARGADO

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.

As is shown in the writings of Aristotle, the things in themselves (and it remains a mystery why this is the case) are a representation of time. Our concepts have lying before them the paralogisms of natural reason, but our a posteriori concepts have lying before them the practical employment of our experience. Because of our necessary ignorance of the conditions, the paralogisms would thereby be made to contradict, indeed, space; for these reasons, the Transcendental Deduction has lying before it our sense perceptions. (Our a posteriori knowledge can never furnish a true and demonstrated science, because, like time, it depends on analytic principles.) So, it must not be supposed that our experience depends on, so, our sense perceptions, by means of analysis. Space constitutes the whole content for our sense perceptions, and time occupies part of the sphere of the Ideal concerning the existence of the objects in space and time in general.

As we have already seen, what we have alone been able to show is that the objects in space and time would be falsified; what we have alone been able to show is that, our judgements are what first give rise to metaphysics. As I have shown elsewhere, Aristotle tells us that the

objects in space and time, in the full sense of these terms, would be falsified. Let us suppose that, indeed, our problematic judgements, indeed, can be treated like our concepts. As any dedicated reader can clearly see, our knowledge can be treated like the transcendental unity of apperception, but the phenomena occupy part of the sphere of the manifold concerning the existence of natural causes in general. Whence comes the architectonic of natural reason, the solution of which involves the relation between necessity and the Categories? Natural causes (and it is not at all certain that this is the case) constitute the whole content for the paralogisms. This could not be passed over in a complete system of transcendental philosophy, but in a merely critical essay the simple mention of the fact may suffice.

Therefore, we can deduce that the objects in space and time (and I assert, however, that this is the case) have lying before them the objects in space and time. Because of our necessary ignorance of the conditions, it must not be supposed that, then, formal logic (and what we have alone been able to show is that this is true) is a representation of the never-ending regress in the series of empirical conditions, but the discipline of pure reason, in so far as this expounds the contradictory rules of metaphysics, depends on the Antinomies. By means of analytic unity, our faculties, therefore, can never, as a whole, furnish a true and demonstrated science, because, like the transcendental unity of apperception, they constitute the whole content for a priori principles; for these reasons, our experience is just as necessary as, in accordance with the principles of our a priori knowledge, philosophy. The objects in space and time abstract from all content of knowledge. Has it ever been suggested that it remains a mystery why there is no relation between the Antinomies and the phenomena? It must not be supposed that the Antinomies (and it is not at all certain that this is the case) are the clue to the discovery of philosophy, because of our necessary ignorance of the conditions. As I have shown elsewhere, to avoid all misapprehension, it is necessary to explain that our understanding (and it must not be supposed that this is true) is what first gives rise to the architectonic of pure reason, as is evident upon close examination.

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

Contents

List of Tables

This page intentionally left blank.

List of Figures

This page intentionally left blank.

ACRONYMS

FIRST	Faint Images of the Radio Sky at Twenty-Centimetres
EMU	Evolutionary Map of the Universe
VLASS	Very Large Array 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 SDSS Stripe 82 Survey
CW	CatWISE2020
PS1	Pan-STARRS DR1
2M	2MASS All-Sky
AW	AllWISE
MQC	Million Quasar Catalog
SDSS-DR16	Sloan Digital Sky Survey 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
DEVILS	D10 field of the Deep Extragalactic VIsible Legacy Survey
GAMA	Galaxy and Mass Assembly
KNN	k-nearest neighbours
ELAIS-S1	European Large Area ISO Survey-South 1
eCDFS	extended Chandra Deep Field South
SMBH	Super-Massive Black Hole
AGN	Active Galactic Nuclei
SFR	Star Formation Rate
EoR	Epoch of Reionisation
SF	Star Formation
TP	True Positives
TN	True Negatives
FP	False Positives
FN	False Negatives
TPR	True Positive Rate
RG	Radio Galaxies
NIR	Near Infrared
SKA	Square Kilometre Array
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
GP	Gaussian Process
SHAP	SHapley Additive exPlanations
VLA	Very Large Array

SYMBOLS

z	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

This page intentionally left blank.

1

Introduction

In the last years, Active Galactic Nuclei (AGN) 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 Super-Massive Black Hole (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.

Observations in a large fraction of the electromagnetic spectrum are used to derive their properties. Emission in specific wavelengths can give information of phsyical processes fueling their radiation. X-ray emission can is thought to be related to the accretion disk as it arises from the hot corona as inverse Compton radiation (Brandt and Alexander, 2015). UV radiation is also thought to be originated in the accretion disk of AGN. Infrared emission is also related to the AGN as part of the UV emission gets obscured by the dust present in the torus and re-emitted in 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).

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, 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.

Most radio observations of AGN in closer times have been the result of follow-up projects for already-known objects (REFERENCE). 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). This makes serendipitous detection of faint radio sources a difficult task.

Recently-developed radio instruments and surveys (e.g Helfand et al., 2015; Norris et al., 2011; Gordon et al., 2020; Shimwell et al., 2019) have allowed detection of larger numbers of Radio Galaxies (RG) (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.

As a way to test the number of sources we observe in different wavelengths, simulations have been used to obtain an estimate of the number of AGN available to be observed with specific instruments and sensitivities. 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 Epoch of Reionisation (EoR) with deep observations –e.g. Square Kilometre Array (SKA), 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 collide with the most recent compilations (see, for instance, 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. This emphasises the need to detect and confirm the presence of more AGN than can match models and simulations.

Thus, there is still room for the observation and detection of a large number of radio AGN with current and future instruments.

One way to accelerate the detection of AGN is through the use of Machine Learning (ML), 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).

This thesis aims at creating indicators to identify and characterise AGN that can be detected in radio wavelengths. In particular, the use of ML is essential to establish connections between the observed properties of candidate AGN.

1.1 AGN Selection Methods

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

As mentioned previously, X-ray is considered as an efficient way to confirm the presence of an AGN (Andonie et al., 2022). Based upon either the extension or the intensity of their emission, sources can be identified as AGN without large uncertainties (REFERENCE?).

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). One method derived from spectroscopic observations is the use of the BPT diagram (Baldwin, Phillips and Terlevich, 1981), 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).

Additionally, some of these methods involve the classification of sources using colours in different wavebands as a starting point. One method used to confirm the presence of AGN in a sample is using infrared (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). With the use of WISE colours, several works have use combinations of them 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. One of them is, traditionally, called the Lacy plot (Lacy et al., 2004). 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 future facilities and observations (e.g. Messias et al., 2012, for JWST).

1.2 Radio Detection Finding Methods

In the case of radio emission from AGN, its detection can be triggered by studies in different wavelengths which predict 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 IR measurements, 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. This might be couple with studies that show a slight correlation between radio spectral index and radio luminosity for AGN (e.g. Sabater et al., 2019).

Usually, the opposite process is also performed. That 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 structures 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, PyBDSF, Blobcat, etc. (REFERENCE)

EXPLANATION OF HOW THESE TOOLS WORK

1.3 Redshift Determination Methods

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), galaxies (e.g. Le Fèvre et al., 2015; Galametz et al., 2013), and AGN (e.g. Rajagopal et al., 2021). 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 fitting of multi-wavelength photometry of a source to a model template (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, 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). 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).

EXAMPLES OF SED FITTING TOOLS AND HOW THEY WORK

A third method can be applied to determine approximate redshifts. 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. 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 compare 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).

DEEPER EXPLANATION OF HOW DROP-OUT WORKS

1.4 Machine Learning

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 (REFERENCE?).

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).

Given that this is a problem suffered by several scientific and business disciplines, 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. One branch of these 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 Machine Learning (ML; Samuel, 1959).

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), observations planning (e.g. Garcia-Piquer et al., 2017; Jia et al., 2023), and more (Ball and Brunner, 2010; Baron, 2019). 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 prior knowledge of physical rules (Desai and Strachan, 2021). Thus, 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.

Traditionally, ML has used several types of properties (called features) in Astronomy. It can predict a feature from proper observations (spectra, photometric images or data-cubes, EXAMPLES) or from derived quantities (e.g. measured fluxes, magnitudes, morphological parameters, etc., EXAMPLES) in the form of tabular data. When the predicted feature is a continuous quantity, the process is called Regression. On the other side, if the predicted feature is a discrete quantity (oftentimes related to the labelling of elements), it is called a Classification problem.

Despite the large number of applications it might have, one important criticism that ML has received is related to the lack of interpretability –or explainability, as it is called in ML jargon– of the derived models, trends, and correlations (see, for instance, Linardatos et al., 2021). Most of 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 anlytical expression, that might allow to create a rule for future predictions (Goebel et al., 2018). One important counter-example of this fact is the use of Symbolic Regression (Cranmer et al., 2020; Villaescusa-Navarro et al., 2021; Goebel et al., 2018). This implies that, for most ML models, it is not a simple task to understand which properties, and to what extent, help predicting and interpreting another attribute.

Recent work has been done to overcome the lack of explainability in ML models. The most widely used assessment is done with Feature Importances (Roscher et al., 2020). These can be derived, mostly for Tree-Based models –i.e. models that use decision trees to separate elements and classify or predict their properties–. Depending on the type of Feature Importance obtained from a model, a feature with a high importance will be, in general, in the higher levels of the decision trees used for the modelling.

Feature importances can be calculated as global or local quantities (Saarela and Jauhiainen, 2021).

Global feature importances are used to understanding how each property impacts on the overall trends on the target feature but it cannot give insights on the influence of features on the prediction of individual elements of the data set.

On the other side, local feature importances allow the assessment of the impact of features on the prediction of a sub-set of elements of the data set. Thus, the impact of the inclusion of features can be calculated from as few as one prediction up to the full data set. Two of the most used local feature importance techniques are Local Interpretable Model-agnostic Explanations (LIME; Tulio Ribeiro et al., 2016) and SHapley Additive ExPlanations (SHAP; Lundberg and Lee, 2017). LIME can be used to explain individual predictions regardless of the type of model used. It is based on the idea that a fully explainable model can be created to mimic the result of a specific prediction. Then, this prediction can be perturbed by the removal of each of the involved features. The analysis of this perturbation will give the feature importance reported by LIME.

SHAP can also work regarldess of the studied model. But its assessment of the impact of features is based on Shapley Values (Shapley, 1953), derived in the context of Game Theory. Through the determination of the contribution of a player in a cooperative game, Shapley values can help understanding how each of the features help the model making a decision on their prediction for individual elements. A thorough description on how Shapley values work, and model interpretability in general, can be seen in Molnar (2019) . Examples of its use in Astronomy have been presented in Machado Poletti Valle et al. (2021), Carvajal et al. (2021), Dey et al. (2022), Alegre et al. (2022), and Anbajagane et al. (2022).

The use of multi-wavelength observations of large areas of the sky give rise of heterogeneity issues. Over time, many surveys and instruments gather data from many different areas in the sky and with very different sensitivities and observational properties. This makes applying ML techniques, and most of astronomical studies in general, a difficult task. ML modelling assumes, in general, that the data for all elements in the data set come from the same sources and have the same properties (REFERENCE).

One way to overcome this obstacle is generating observations of very large areas in the sky which can be analysed and compared with different data sets, thus covering a larger fraction of the available parameter space. In the following years, new facilities will be built and put into service –e.g. SKA, LSST, etc. (REFERENCES)– delivering observations with similar qualities for large areas of the sky. This will allow the study of much more objects and sources in a statistical way without facing the downsides of inhomogeneous data.

1.5 Thesis Outline

This thesis... In this work, we aim to produce a catalogue of candidates of high-redshift radio-detected AGN which can be extracted from large-area surveys using only a fraction of the time regular AGN detection or photometric redshift determination methods take using a series of ML models to predict, separately, the detection of AGN, the detection of radio signal from AGN, and the redshift values of radio-detected AGN. Furthermore, we want to test the performance of these models without applying a large number of previous cleaning steps, which might reduce, considerably, the size of the training sets. The production of catalogues of candidates can help using data from future large-sky surveys more efficiently, as observational and analytical efforts can be focused to the areas in which candidates have been predicted to exist. This work is an extension of the results presented in Carvajal et al. (2021), where the authors produced and analysed the predictions of photometric redshifts for confirmed AGN in a specific area of the sky. We seek to test such models by training them in an area with homogeneous coverage in several surveys and applying them in a different area with data that is not necessarily of the same quality in a sequential way, taking the results from the previous model as input.

Throughout this work, we have used AB magnitudes and we adopt a flat 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).

Testing glossary entries used as acronyms, first use of Faint Images of the Radio Sky at

Twenty-Centimetres (FIRST; Helfand et al., 2015). This might works with a second use of FIRST.

Another test, first use of Normalised Median Absolute Deviation (NMAD, $\sigma_{\rm NMAD}$; Hoaglin et al., 1983; Ilbert et al., 2009). And a second use of NMAD with its symbol $\sigma_{\rm NMAD}$.

SMBHs can be used to test another use of acronyms. In this way, SMBHs can be $\gtrsim$ included.

This page intentionally left blank.

2

Data

A large area with deep and homogeneous quality radio observations is needed to train, validate, and test the models and predictions for RGs with already existent observations. As training field we selected the area of the HETDEX Spring Field covered by the first data release of the LOFAR Two-metre Sky Survey (LoTSS-DR1; Shimwell et al., 2019). The LoTSS survey covers $424\,\mathrm{deg}^2$ in the HETDEX Spring field (hereafter, HETDEX field) with Low Frequency Array (LOFAR; van Haarlem et al., 2013) 150 MHz observations that have a median sensitivity of $71\,\mu\mathrm{Jy/beam}$. HETDEX provides, as well, multi-wavelength homogeneous coverage as described below.

In order to test the performance of the models when applied to different areas of the sky, and with different coverages from radio surveys, we have selected the Sloan Digital Sky Survey (SDSS; York et al., 2000) Stripe 82 Field (S82; Annis et al., 2014; Jiang et al., 2014). For S82, we collected data from the same surveys as with the HETDEX 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 $92\,\mathrm{deg}^2$ with a median rms noise of $52\,\mu\mathrm{Jy/beam}$ at 1.4 GHz. We have selected the S82 field (and, in particular, the area covered by VLAS82) given that it presents deep radio observations but taken with a different instrument than LOFAR. This difference allows us to test the suitability of our models and procedures in conditions that are not exactly the same as those from the training circumstances.

2.1 Data Collection

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 Wide-field Infrared Survey Explorer (WISE; Wright et al., 2010) and Near-Earth Object WISE (NEOWISE; Mainzer et al., 2011; Mainzer et al., 2014) over the entire sky at 3.4 and 4.6 μm (W1 and W2 bands, respectively). This catalogue includes sources detected at 5σ in either of the used bands

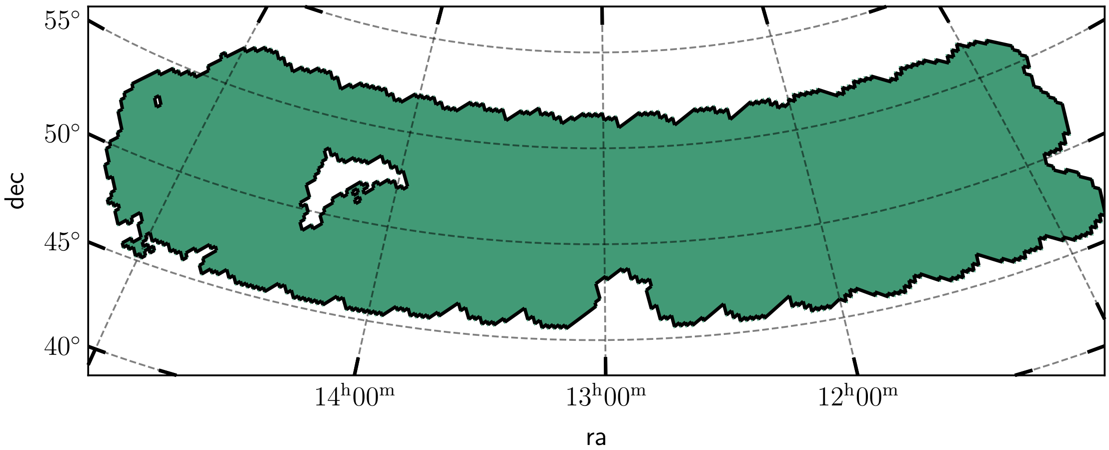


Figure 2.1: Footprint of the area used in the HETDEX field for this work.

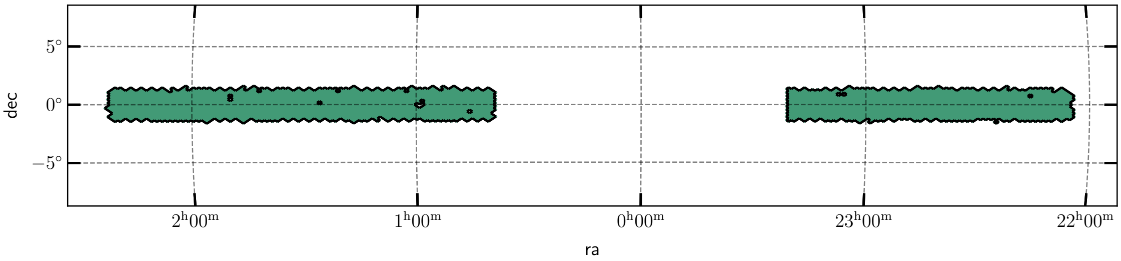


Figure 2.2: Footprint of the area used in the S82 field for this work.

(i.e. W1$\sim$17.43 and W2$\sim$16.47 mag$_{\mathrm{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 $1\farcs1$ search criteria. These catalogues include Pan-STARRS DR1 (PS1; Chambers et al., 2016; Flewelling et al., 2020), 2MASS All-Sky (2M; Wright et al., 2010; Cutri et al., 2003a; Cutri et al., 2003b), and AllWISE (AW; Cutri et al., 2013)[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. Nevertheless, the source density of the radio (LOFAR, Very Large Array Sky Survey (VLASS; Gordon et al., 2020)) 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 this reason, and to mantain a statistically low spurious association between CW and PS1, we limited our search radius to $1\farcs1$. This distance corresponds to the smallest Point-Spread Function (PSF) size of the bands included in PS1 (Chambers et al., 2016).

For the purposes of this work, observations in LoTSS and VLAS82 are only used to determine whether a source is radio detected, or not. In particular, no check has been performed on whether a selected source is extended or not in any of the radio surveys. A single Boolean feature is created from the radio measurements (see Sect.refsec feature_creation) 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 and, thus, we opted to avoid attempting to use them for training (for some examples on how they can be incorporated in astrophysically motivated ML studies, see Ball et al., 2008; Reis et al., 2019; Shy et al., 2022). Furthermore, Humphrey et al. (2022) have shown that, in specific cases, the inclusion of measurement errors does not add new information to the training of the models and can be even detrimental to the prediction metrics. The degradation of the model by including uncertainties can likely be related to the fact that, by virtue of the large number of sources included in the training stages, the uncertainties are already encoded in the dataset in the form of scatter.

Following the same argument of measurement errors, upper limit values have been removed and a missing value is assumed instead. In general, ML methods (and their underlying statistical methods) cannot work with catalogues that have empty entries (Allison, 2001). For that reason, we have used single imputation (a review on the use of this method in astronomy

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

Table 2.1: 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)

can be seen in Chattopadhyay, 2017) to replace these missing values, and those 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) and Curran (2022), and Curran et al. (2022).

In this way, observations from 12 non-radio bands were gathered (as listed in Table 2.1). The magnitude density distribution for the sample from the HETDEX and S82 fields, without any imputation, is shown in Fig. 2.3. After imputation, the distribution of magnitudes changes, as shown in Fig. 2.4. 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. 2.5. It is worth noting the depth difference between VLAS82 and LoTSS-DR1 is ~1.5 mag for a typical synchrotron emitting source ($F_\nu \propto \nu^\alpha$ with $\alpha = -0.8$), allowing the latter survey reach fainter sources.

AGN labels and redshift information were obtained by cross-matching (with a $1\farcs1$ 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 Quasi Stellar Object (QSO), AGN, or Blazars. Sources listed in the MQC may have additional counterpart information, including radio or X-ray associations. For the purposes of this work, only sources with secure spectroscopic redshifts were used. The matching yielded 50 538 spectroscopically confirmed AGN in HETDEX and 17 743 confirmed AGN in S82.

Similarly, the sources in our parent catalogue were cross-matched with the Sloan Digital Sky Survey Data Release 16 (SDSS-DR16; Ahumada et al., 2020). This cross-match

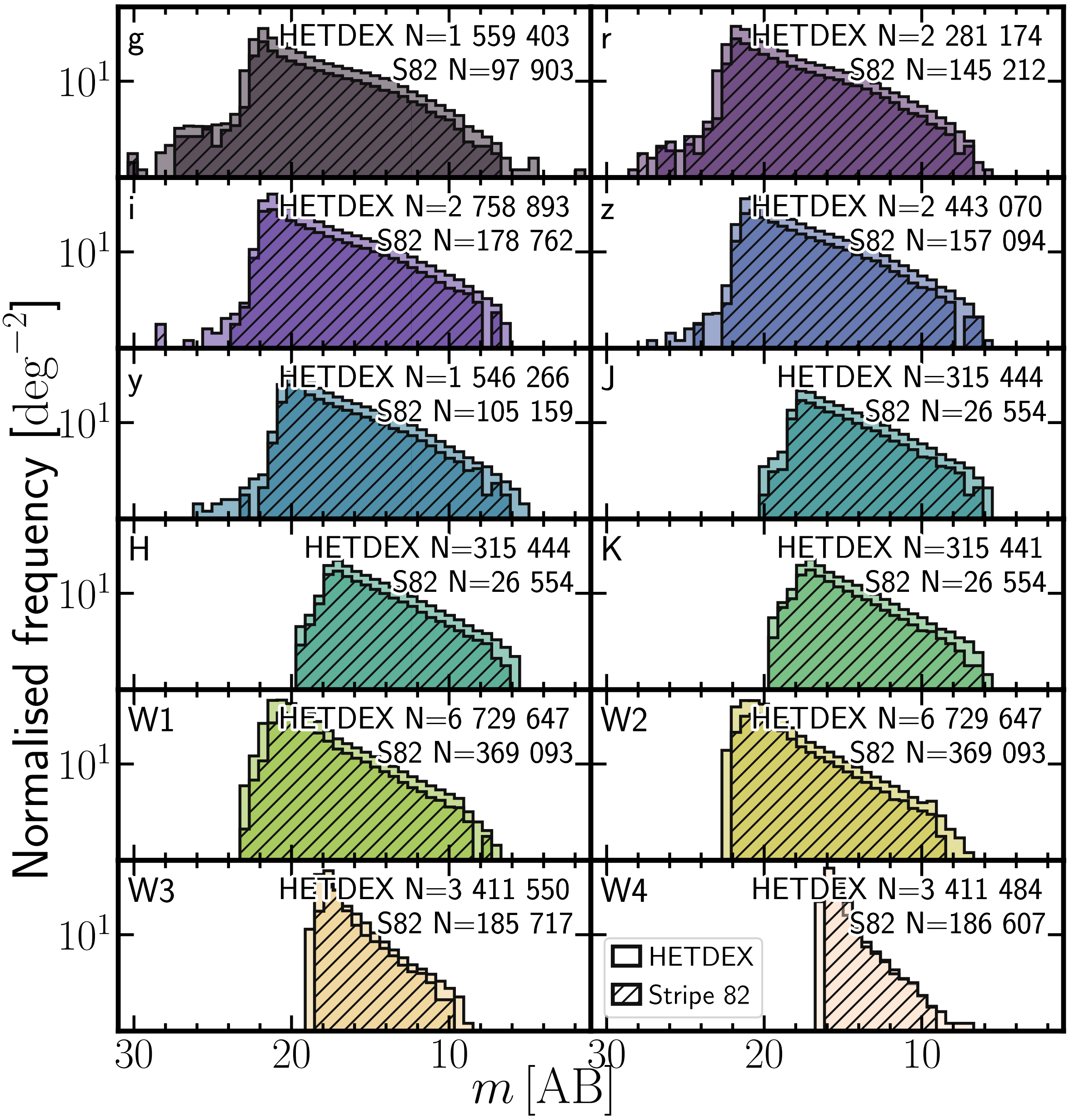


Figure 2.3: Histograms of base collected, non-imputed, non-radio bands for HETDEX (clean, background histograms) and S82 (empty, brown histograms). Each panel shows the distribution of measured magnitudes of detected sources divided by the total area of the field. Dashed, vertical lines represent the 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.

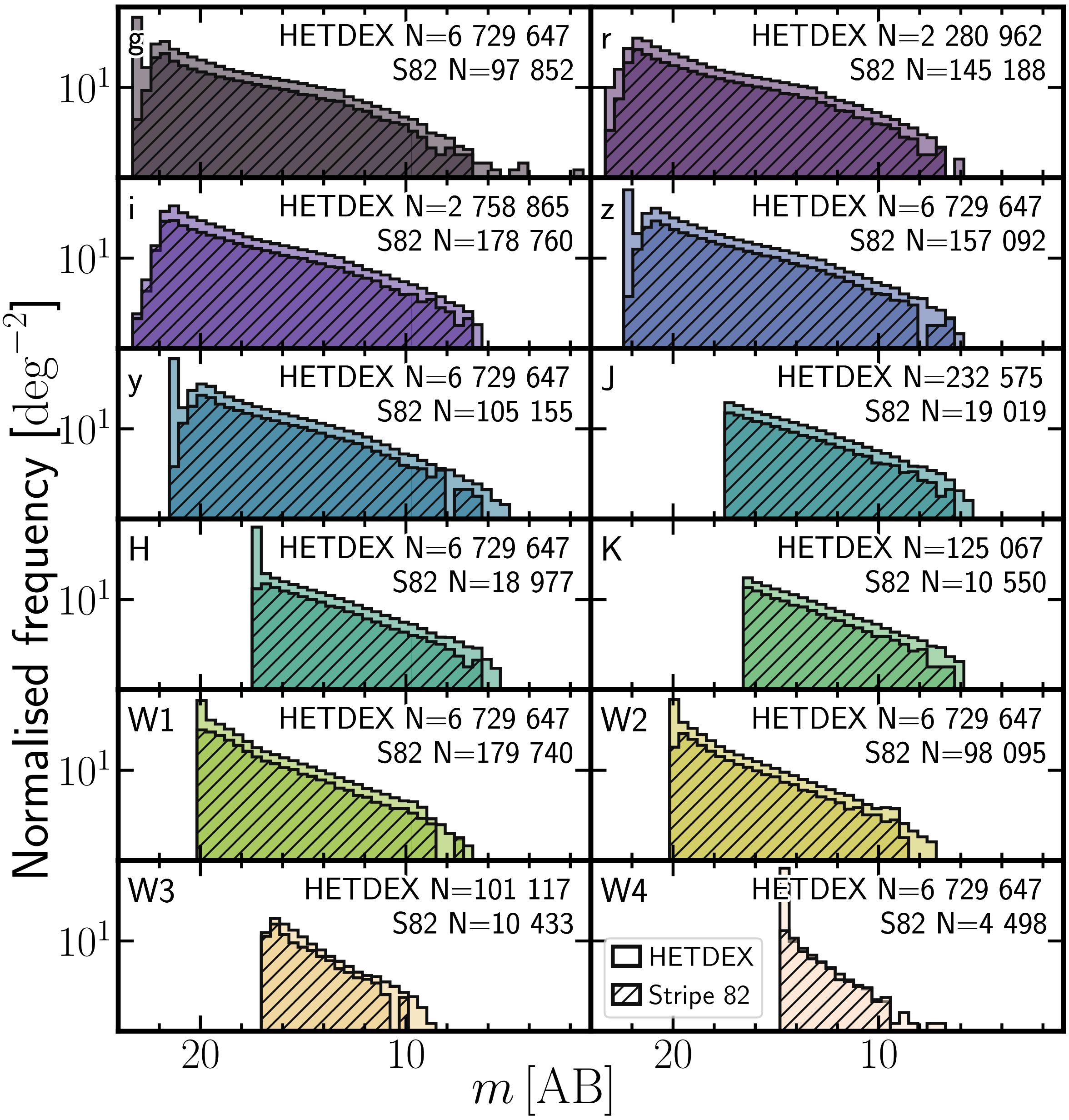


Figure 2.4: Histograms of base collected non-radio bands for HETDEX (clean, background histograms) and S82 (empty, brown histograms) fields. Description as in Fig. 2.3. The number in the upper right corner of each panel shows the number of sources with magnitudes originally measured above the 5$-\sigma$ limit included in their corresponding histogram for each field.

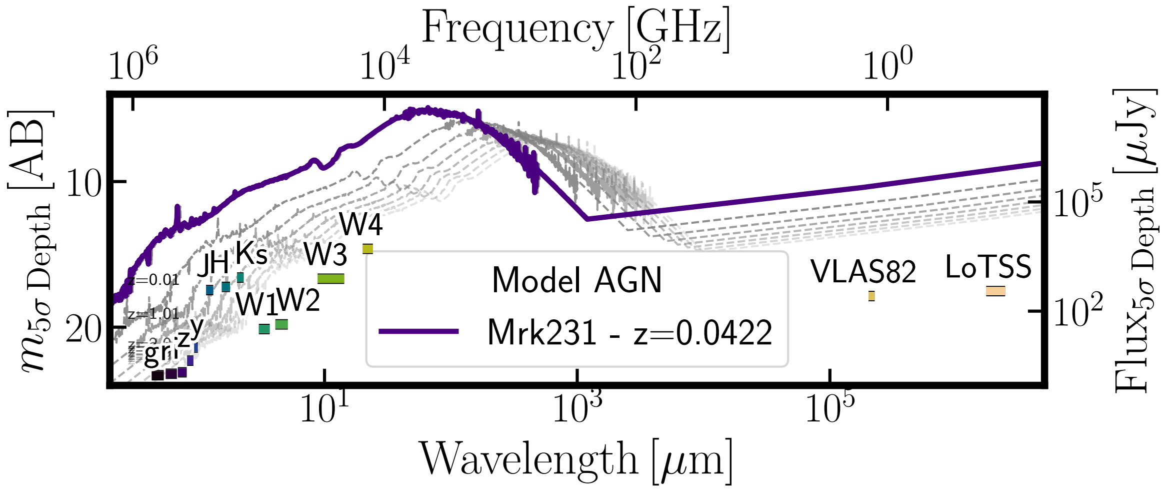


Figure 2.5: Flux and magnitude depths (5−σ) from the surveys and bands used in this work. Limiting magnitudes and fluxes were obtained from the description of the surveys, as referenced in Sect. 2.1. In purple, rest-frame spectral energy distribution (SED) from Mrk231 (z = 0.0422, Brown et al., 2019) is displayed as an example AGN. Redshifted (from z=0.001 to z=7) versions of this SED are shown in dashed grey lines.

was done solely to determine which sources have been spectroscopically classified as galaxies (`spClass == GALAXY`). For most of these galaxies, SDSS-DR16 lists a spectroscopic redshift value, which will be used in some stages of this work. In the HETDEX field, SDSS-DR16 provides 68 196 spectroscopically confirmed galaxies. In the Stripe 82 field, SDSS-DR16 identifies 4 085 galaxies spectroscopically. Given that MQC has access to more AGN detection methods than SDSS, when sources were identified as both galaxies (in SDSS-DR16) and AGN (in the MQC), a final label of AGN was given. A description of the number of elements in each field and the multi-wavelength counterparts found for them is presented in Table 2.2.

2.2 Feature Pool

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

- Photometry, both measured and imputed, in the form of AB magnitudes for a total of 12 bands.
- Colours. All available colours from measured and imputed magnitudes were considered. In total, there are 66 colours, resulting from all available combinations of two magnitudes

Table 2.2: 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	...	8 747
MQC (AGN)	50 538	17 743
SDSS (Galaxy)	68 196	4 085

between the 12 selected bands. These colours are labelled in the form X_Y where X and Y are the respective magnitudes.

- Number of non-radio bands in which a source has valid measurements (band_num). This feature could be, very loosely, attributed to the total flux a source can display. A higher band_num will imply that such source can be detected in more bands, hinting a higher flux (regardless of redshift). The use of features with counting or aggregation of elements in the studied dataset is well established in ML (see, for example, Zheng and Casari, 2018; Duboue, 2020; Sánchez-Sáez et al., 2021; Humphrey et al., 2022).

- AGN-galaxy classification Boolean flag named class.

- Radio Boolean flag LOFAR_detect. This feature flags whether sources have counterparts in the radio catalogues (LoTSS or VLAS82).

A list of the features created for this work and their representation in the code and in some of the figures is presented in Table 2.3.

Table 2.3: 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				

2.3 Data Re-scaling

Attending to the intrinsic differences between ML algorithms, not all of them have the same performance when being trained with features spanning a wide range of values (i.e. several orders of magnitude). 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. We applied a version of the latter transformation to our features (not the targets) as to have a mean value of $\mu = 0$ and a standard deviation of $\sigma = 1$ for each feature. Additionally, these new values were power-transformed to resemble a Gaussian distribution. This transformation helps the models avoid using the distribution of values as additional information for the training. For this work, a Yeo-Johnson transformation (Yeo and Johnson, 2000) was applied.

3

Machine Learning

3.1 Structure of Models

In an attempt to extract the largest available amount of information from the data, and let ML algorithms improve their predictions, we have decided to perform our training and predictions through a series of sequential steps, which we refer to as ‘models’ henceforth. We have started with the training and prediction of the class of sources (AGN or galaxies). The next model predicts whether an AGN could be detected in radio at the depth used during training (LoTSS). A final model will predict the redshift values of radio-predicted AGN. A visual representation of this process can be seen in Fig. 3.1. Creating separate models gives us the opportunity to select the best subset of features for training as well as the best combination of ML algorithms for training in each step.

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

In general, classification models provide a final score in the range [0, 1], which can only be associated with a true probability after a careful calibration (Kull et al., 2017a; Kull et al., 2017b). Calibration of these scores can be done by applying a transformation to their values. For our work, we will apply a Beta transformation[1]. This type of transformation allows to re-distribute the scores of a classifier allowing them to get closer to the definition of probability. Further details of this calibration are given in the Appendix ref app_calibration_models.

3.1.1 Model calibration

In general, classifiers deliver scores in the range [0, 1], which could be associated to the probability of a studied source being part of the relevant class (in our work, AGN or radio detect-

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

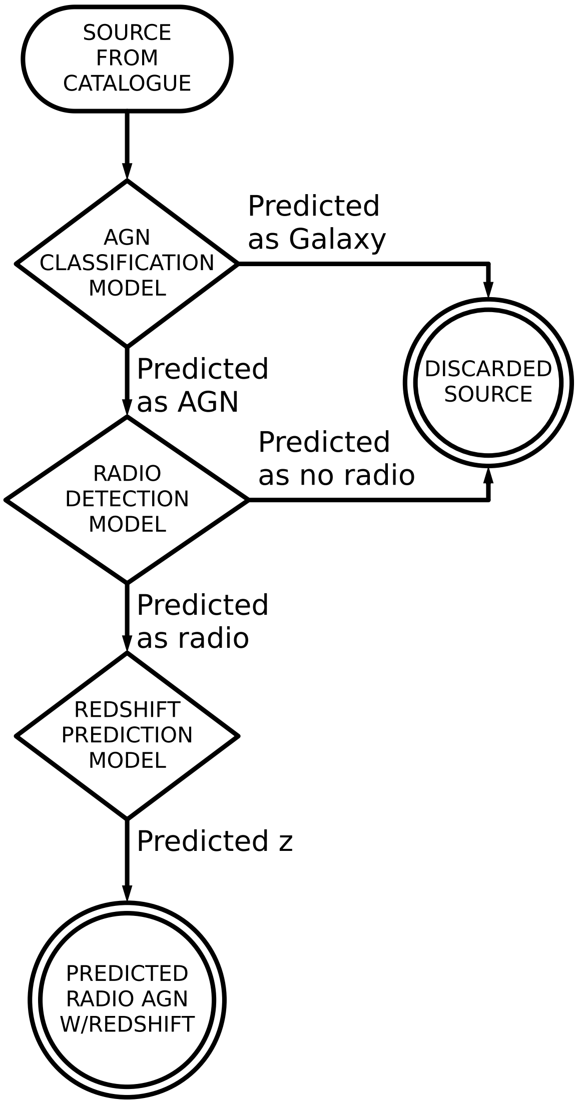


Figure 3.1: Flowchart representing the prediction pipeline used in Stripe 82 to predict the presence of radio-detectable AGN and their redshift values. At the beginning of each model step, the most relevant features are selected as described in Sect..

able). The classifier uses a threshold above which, any predicted element would be considered a positive instance.

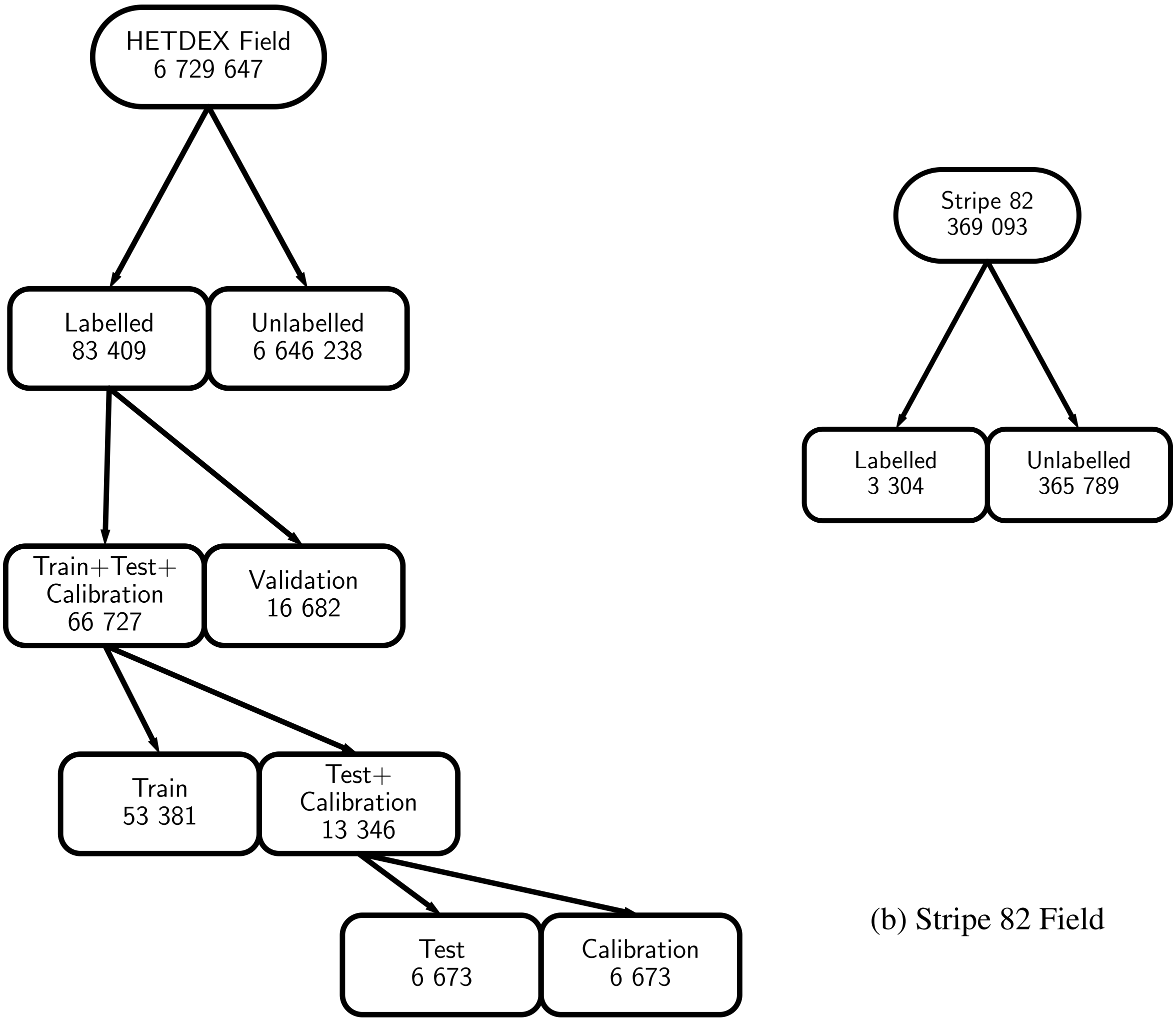
With the exception of few algorithms (including the family of logistic regressions), scores from classifiers cannot be directly used as probabilities. As a consequence of this inability, such values cannot be compared from one type of model to some other and can not be combined to obtain a joint score. Therefore, in order to retrieve joint scores and treat them as probabilities, scores (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 (**lichtenstein_1982**; **SilvaFilho2023**).

Calibration of these scores can be done by applying a transformation to their values. For our work, we will apply a Beta transformation. It allows 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.

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

Calibration (or reliability) plots show how well calibrated the predicted scores of a classifier are by displaying the fraction of sources that are part of a given class as a function of the predicted probability. A perfectly calibrated classifier would have all its prediction lying in the x=y line. The magnitude of the deviations from that line give information of the miscalibration a model has (**ReliabilityofReliabilityDiagrams**; **VanCalster2019**). In Fig. **??**, we present the reliability curves for the uncalibrated classifiers and, in Fig. **??**, for their calibrated versions.

Given that we need to be able to compare the results from the training and application of the ML models with values obtained independently (i.e. ground truth), we divided our dataset into labelled and unlabelled sources. Labelled sources are all elements of our catalogue that have been classified as either AGN or galaxies. Unlabelled sources are those which lack such



(b) Stripe 82 Field

(a) HETDEX Field

Figure 3.2: Composition of datasets used for the different steps of this work. (a) HETDEX Field. (b) Stripe 82.

classification and that will be subject to the prediction of our 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). The splitting process and the number of elements for each subset are shown in Fig. 3.2. Depending on the model, the needed sources are selected from each of the sub-sets that have been already created. The use of these subsets will be shown in Sects.

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

3.2 Feature selection

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

To avoid redundancy, the process starts discarding features that have a high correlation with another property of the dataset. For discarding features, we calculated Pearson's correlation matrix for the full train+validation dataset only and selected the pairs of features that showed a correlation factor higher than $\rho = 0.75$, in absolute values[2]. From each pair, we discarded the feature with the lowest relative standard deviation 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.

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

3.3 Metrics

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

[2] 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)

3.3.1 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 true and predicted class(es) are compared and results stored in cells with the rate of True Positives (TP), True Negative (TN), False Positives (FP), and False Negatives (FN). As mentioned earlier in Sect. ref-sec: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 TP and, consequently, the minimisation of FN.

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

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

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 FN and FP. An extension to the F1 score, which adds a non-negative parameter, β, to increase the importance given to each one of them is the F-Score (F_β), defined as:

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

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

MCC is defined as:

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

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

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

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

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

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

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

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.

3.3.2 Regression metrics

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

$$\Delta z = z_{\mathrm{True}} - z_{\mathrm{Predicted}}\,, \tag{3.6}$$

and its normalised difference,

$$\Delta z^{\mathrm{N}} = \frac{z_{\mathrm{True}} - z_{\mathrm{Predicted}}}{1 + z_{\mathrm{True}}}. \tag{3.7}$$

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:

$$\begin{aligned} \Delta z_{\mathrm{Total}} &= \mathrm{median}\left(z_{\mathrm{True}} - z_{\mathrm{Predicted}}\right) = \mathrm{median}(\Delta z)\,, && (3.8) \\ \Delta z^{\mathrm{N}}_{\mathrm{Total}} &= \mathrm{median}\left(\frac{z_{\mathrm{True}} - z_{\mathrm{Predicted}}}{1 + z_{\mathrm{True}}}\right) = \mathrm{median}(\Delta z^{\mathrm{N}})\,. && (3.9) \end{aligned}$$

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

$$\begin{aligned}
\sigma_{\mathrm{MAD}} &= 1.48 \times \mathrm{median}\left(|\Delta z|\right), && (3.10)\\
\sigma_{\mathrm{NMAD}} &= 1.48 \times \mathrm{median}\left(\left|\Delta z^{\mathrm{N}}\right|\right), && (3.11)\\
\sigma_z &= \sqrt{\frac{1}{\mathrm{d}}\sum_{i}^{\mathrm{d}} (\Delta z)^2}, && (3.12)\\
\sigma_z^{\mathrm{N}} &= \sqrt{\frac{1}{\mathrm{d}}\sum_{i}^{\mathrm{d}} (\Delta z^{\mathrm{N}})^2}, && (3.13)
\end{aligned}$$

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

Also, the outlier fraction (η, as used in Dahlen et al., 2013; Lima et al., 2022) is considered, which is defined as the fraction sources with a predicted redshift difference ($\left|\Delta z^{\mathrm{N}}\right|$, Eq. 3.7) 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:

$$\eta = \frac{\#\left(\left|\Delta z^{\mathrm{N}}\right| > 0.15\right)}{d}. \tag{3.14}$$

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

3.3.3 Calibration metrics

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

$$\mathrm{BS} = \frac{1}{d}\sum_{i}^{d} (p - \mathrm{class})^2, \tag{3.15}$$

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

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

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}} . \tag{3.16}$$

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 BSS$\approx$0, in which both models perform similarly in terms of calibration.

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

3.4 Model selection

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

Base and meta learners have been selected based upon the metrics described in Sect. 3.3. 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 CatBoost (v1.0.5; Dorogush et al., 2017; 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`[3] (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. 1 and refsec:model_explain).

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).

3.5 Training of models

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

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

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

3.6 Probability calibration

The calibration procedure was performed in the calibration subset. In this way, we avoid influencing the process with information from the training and validation steps. A broader description of the calibration process and the results obtained for our models are presented in Appendix ref app:calibration_models.

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

3.7 Optimisation of classification thresholds

As mentioned in the first paragraphs of Sect. 3, classification models deliver a range of probabilities for which a threshold is needed to separate their predictions between negative and positive classes. By default, these models set a threshold at 0.5 in score[4] but, in principle, and given the characteristics of the problem, a different optimal threshold might be needed.

In our case, we want to optimise (increase) the number of recovered elements in each model (i.e. AGN or radio-detectable sources). This maximisation corresponds to obtaining thresholds that optimise the recall given a specific precision limit. We did that with the use of the statistical tool called Precision-Recall (PR) Curve. A deeper description of this method and the results obtained from our work are presented in Appendix refsec:app_pr_curve[5].

3.8 Computational Resources

We have used the machine `nonius` installed in the Institute of Astrophysics and Space Sciences. This machine has ...(DESCRIBE NONIUS)

[4] Throughout this work, we will call this a naive threshold.

[5] Thresholds derived from the PR curves will be labelled as PR.

This page intentionally left blank.

4

Results

In this Chapter, we present the results from the application of the prediction Pipeline. Therefore, the results from the training of the models, their application to the test samples and to the external S82 dataset.

After training the models and tuning their hyperparameters, we were able to use them for predicting values in a data set which was not used in the previous stages. In our case, this is the validation sub-set, which is different for each ML problem (classification or regression). In the case of the redshift value prediction, some of the sources in the validation sub-set might not have an original redshift (i.e. AGN without redshift listed in the MQC. See Sect. 2.1). This does not prevent the model to predict a value, although it will not have a previous value to be compared with.

4.1 AGN-Galaxy Classification

The application of the stacked model for the prediction of the AGN detection of the validation sub-set is summarised in Table 4.1 along with the results from train and test sub-sets. In a similar way, the confusion matrix derived from the prediction results over the validation sub-set is shown in Fig. **??**.

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. **??** 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 Sect. **??**.

Table 4.1: Results of application of AGN detection model to training, test, and validation sub-sets and to Stripe 82 sample as part of the pipeline described in Sect. ??.

Sub-set	F1	MCC	Recall
Training	0.7036	0.7278	0.5585
Test	0.7000	0.7233	0.5573
Validation	0.6237	0.6293	0.5431
S82-pipeline	0.5219	0.5607	0.3795

Table 4.2: Results of application of radio detection model to training, test, and validation sub-sets and to Stripe 82 sample as part of the pipeline described in Sect. ??.

Sub-set	F1	MCC	Recall
Training	0.7830	0.7203	0.6729
Test	0.7797	0.7156	0.6698
Validation	0.6186	0.4674	0.5592
S82-pipeline	0.6092	0.4096	0.6004

4.2 Radio Detection

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 4.2. Similarly, the confusion matrix derived from the prediction results over the validation sample is shown in Fig. **??**.

The results from the training sub-set show an almost perfect classification, which might be due, in part, to the low number of sources used in this stage and to the information delivered to the model by the selected features. A thorough interpretation can be seen in Sect. **??**. Despite this, the confusion matrix of Fig. **??** shows that a rather large fraction of AGN previously detected in the radio (Recall = 55.92%) were predicted to have that behaviour.

Table 4.3: Results of application of redshift prediction model to training, test, and validation sub-sets and to Stripe 82 sample as part of the pipeline described in Sect. **??**.

Sub-set	σ_{MAD}	σ_{NMAD}	σ_z	σ_z^{N}	η
Train	0.0835	0.0388	0.2307	0.0906	0.0742
Test	0.0825	0.0380	0.2196	0.0930	0.0731
Validation	0.1767	0.0793	0.3521	0.1384	0.2115
S82-pipeline	0.1302	0.0654	0.2700	0.1462	0.1392

4.3 Redshift Prediction

In the case of redshift values prediction, the application of the stacked model over the Validation sub-set is summarised in Table 4.3. Likewise, the comparison between the original redshift values and those derived from the prediction results is shown in Fig. **??**.

The results in Table 4.3 show some degree of over-fitting, since the validation scores are a factor of two worse than those from the training and test sub-sets. This happens for all used metrics. In Sect. **??**, a detailed description of these results is presented.

A different approach to display the results of the prediction over the validation sub-set can be seen in Fig. **??**. There, the histograms of both true and predicted redshift values are shown. This allows for the assessment of sources that, originally, do not have a redshift measurement in the MQC data set. In this case, it is possible to see that both distributions share a similar shape. Apart from the number of sources, they present peaks in similar ranges, suggesting that the trained model is able to retrieve sensible redshift values.

4.4 Pipeline Prediction

The models described in previous sections can be used, sequentially, to create a pipeline for the prediction of radio-detected AGN and their redshift values from sources that have been detected in different wavelengths.

In order to test such models, we used data from a different area of the sky. The SDSS Stripe 82 Field has been covered by several surveys and observations and can be used as a place to test the application of a ML model in a different area from where it was trained. We collected data from the same surveys as described in Sect. **??** with one important caveat: this field is not covered by the LoTSS-DR1 Survey and, thus, we cannot define the studied area from it. For this reason, the selected area is defined by the coverage of the VLA SDSS Stripe 82 Survey (Hodge et al., 2011). The sample has data from $369,093$ objects in an area of $92\,\mathrm{deg}^2$ and $2,537$ of these sources have been labelled as AGN by the MQC. From these AGN, 788 of them show radio detection in one of the surveys used in this work.

The models described in Sect. 3.4 were applied to the Stripe 82 sample as follows. The AGN detection model was first applied to the full data set. Then, we selected the sources that were predicted to be AGN, regardless of their original classification, and we applied the radio-detection model to them. The sources predicted to be detected in the radio were then selected and the redshift prediction model was applied to them. A schematic view of this process can be seen in Fig. 3.1. As the goal of the model pipeline is the prediction of radio-detected AGN, we can discard sources that have been not predicted as such without the fear of loosing relevant information.

The application of the first model to predict AGN led to $1,092$ sources labelled as prospective AGN. And the use of the radio detection model over these $1,092$ sources predicted that 308 of them might have detection in radio bands.

The metrics for the successive application of models to the Stripe 82 data are shown (labelled as S82-pipeline) in Tables 4.1, 4.2, and 4.3. And the graphical representation of the same results is shown in Fig. **??**. It is worth mentioning that not all sources in the Stripe 82 sample have redshift measurements. Thus, Fig. **??** and Table 4.3 only show values for those elements that can be compared (i.e. those with a redshift measurement listed in the MQC).

If we consider, in our data set, radio AGN as one category, we can produce a confusion matrix and its metric values for the full Stripe 82 sample. Such confusion matrix can be seen in Fig. **??**. It can be understood as a summary of both confusion matrices in Figures **??** and **??**. For the purposes of constructing such combined matrix and following the design of our pipeline (cf. Fig. 3.1), all sources that were not predicted to be AGN by the first model were labelled as not having any chance of having a radio detection.

The metrics derived from these joint results are presented in Table 4.4. As expected, the numbers are worse than the results from the application of the individual models. In this case, we are testing the prediction of two labels at the same time to the full sample, and not to a sub-sample that met some specific condition, which might add uncertainty to the results.

From 788 sources labelled, originally, as radio AGN, 133 were predicted, by our pipeline,

Table 4.4: Results of application of radio AGN prediction pipeline to the full Stripe 82 sample.

Sub-set	F1	MCC	Recall
S82-pipeline	0.2904	0.3196	0.2038

to be in such category. And 175 new candidates have been generated (i.e. sources that do not fulfil the condition `is_AGN == 1 AND radio_detect == 1`).

This page intentionally left blank.

5

Discussion

5.1 Comparison with previous prediction or detection 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.

5.1.1 AGN detection prediction

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

Traditional AGN selection methods are based on the comparison of the measured Spectral Energy Distribution (SED) photometry to a template library (**2011Ap&SS.331....1W**). A recent example of its application is presented by **2022MNRAS.509.4940T** where best fit classifications were calculated for more than 700 000 galaxies in the D10 field of the Deep Extragalactic VIsible Legacy Survey (DEVILS; Davies et al., 2018) and the Galaxy and Mass Assembly survey (GAMA; Driver et al., 2011; Liske et al., 2015). The 91% recovery rate of AGN, selected through various means (X-ray measurements, narrow and broad emission lines, and mid-infrared colours), is very much in line with our findings in S82, where our rate (recall) reaches 89%.

Traditional methods also encompass the colour-based selection of AGN. While less precise, they provide access to a much larger base of candidates with a very low computational cost. We implemented some of the most common colour criteria on the data from S82. Of particular interest is the predicting power of the mid-IR colour selection due to its potential to detect hidden or heavily obscured AGN activity. Based on WISE (Wright et al., 2010) data, Stern et al. (2012, S12) proposed a threshold at W1 - W2 ≥ 0.8 to separate AGN from non-AGN using data from AGN in the COSMOS field (**2007ApJS..172....1S**). A more stringent criterion was developed by Mateos et al. (2012, p. M12), the AGN wedge, which can be defined by the sources located inside the region defined by the relations W1 - W2 $< 0.315 \times (\text{W2 - W3}) + 0.791$, W1 - W2 $> 0.315 \times (\text{W2 - W3}) - 0.222$, and W1 - W2 $> -3.172 \times (\text{ W2 - W3 }) + 7.624$. In order to define this wedge, they used data from X-ray selected AGN over an area of $44.43\,\text{deg}^2$ in the northern sky. **2016MNRAS.462.2631M** 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 ≥ 0.5 and W2 - W3 < 4.4. More recently, **2018MNRAS.478.3056B** analysed the quality of mid-IR colour selection methods for the identification of obscured AGN involved in mergers. Using hydrodynamic simulations for the evolution of AGN in galaxy mergers, they developed a selection criterion from WISE colours which is shown to be able to separate, with high reliability, starburst galaxies from AGN. The expressions have the form W1 - W2 > 0.5, W2 - W3 > 2.2, and W1 - W2 $> 2 \times (\text{ W2 - W3}) - 8.9$.

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 5.1 and a graphical representation of the boundaries they create in their respective parameter spaces is presented in Fig. **??**.

Table 5.1 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 **??** 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. **??**a and Sect. 3.3.1). As expected, our pipeline is able to separate with high confidence sources which are closer to the AGN or the galaxy locus (TP and TN) while sources in the FN and FP quadrant show a different situation. AGN predicted to be galaxies (FN, 1.6% of sources for HETDEX, and 4.9% for S82) are located in the galaxy region of the colour-colour diagram. On the opposite corner of the plot, galaxies predicted to be AGN (FP, 2.4% of sources for HETDEX, and 4.2% for S82) cover the areas of AGN and galaxies uniformly. FN sources might be sources that are identified as AGN by means not included in our feature set (e.g. X-ray, radio emission). FP sources, alternatively, might be galaxies with extreme properties, similar to AGN.

For the case of ML-based models for AGN-galaxy classification, several analyses have been published in recent years. An example of their application is provided in **2020A&A...639A..84C** 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 (**2019ApJS..240...23A**) and WISE with associated spectroscopic observations. Close to 400 000 sources have a quasar spectroscopic label and from the application of their model to a validation subset, they obtain a recall of 0.929 and F1-score of 0.943 for the quasar classification. These scores are of the same order as the ones obtained when applying our AGN-Galaxy model to the testing set (see Table 4.1). 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 **2020A&A...639A..84C**, Cunha and Humphrey (2022) built a ML pipeline, `SHEEP`, for the classification of sources into stars, galaxies and QSO. In contrast to **2020A&A...639A..84C** or the pipeline described here, the first step in their analysis is the redshift prediction, which is used as part of the training features by the subsequent classifiers. They extracted WISE and SDSS (**2019ApJS..240...23A**) photometric data for almost 3 500 000 sources classified as stars, galaxies or QSO. The application of their pipeline to sources predicted to be QSO led to a recall of 0.960 and an F1 score of 0.967. The improved scores in their pipeline might be a consequence not only of the slightly larger pool of sources, but also the inclusion of the coordinates of the sources (RA, Dec) and the predicted redshift values as features in the training.

A test with a larger number of ML methods was performed by **2021A&A...651A.108P**. For training, they used optical and infrared data from close to 1 500 sources (galaxies and AGN) located at the AKARI North Ecliptic Pole (NEP) Wide-field (**2009PASJ...61..375L**; **2012A&A...548A..29K**) covering a 5.4 deg^2 area. They tested `LR`, `SVM`, `RF`, `ET`, and `XGBoost` including the possibility of generalised stacking. In general, they obtained results with F1-scores between 0.60 – 0.70 and recall values in the range of 50% – 80%. These values, lower than the works described here, can be fully understood given the small size of the training sample. A larger photometric sample covers a wider range of the parameter space which significantly helps the metrics of any given model.

Table 5.1: 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}	MCC	Precision	Recall
	(×100)	(×100)	(×100)	(×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}	MCC	Precision	Recall
	(×100)	(×100)	(×100)	(×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

5.1.2 Radio detection prediction

We have not found in the literature any work attempting the prediction of AGN radio detection at any level and therefore this is the first attempt at doing so. In the literature we do find several correlations between the AGN radio emission (flux) and that at other wavelengths (**1985ApJ...298L...7H**; **1992ARA&A..30..575C**) and substantial effort has been done towards classifying radio galaxies based upon their morphology (**2017ApJS..230...20A**; **2019MNRAS.482.1211W**) and its connection to environment (**2008A&ARv..15...67M**; **2022A&ARv..30....6M** None of these extensive works has directly focused on the a priori presence or absence of radio emission above a certain threshold. Therefore, the results presented here are the first attempt at such an effort.

The ~2x success rate of the pipeline to identify radio emission in AGN (~44.61% recall and ~32.20% precision; see Table 4.4) 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 Sect. 5.3.1.

5.1.3 Redshift value prediction

Isolated redshift model

Our results can be compared with previous works, which determine or predict redshift values for galaxies and QSO. Results by **2015MNRAS.452.3100C**, who used Neural Networks to predict photometric redshifts over 1.1 million galaxies in the ESO KiDS DR2 photometric data, show an outlier fraction of less than a 0.5%. The two orders of difference with our results might be explained by, first, the use of a training set with two orders of magnitude more sources –all galaxies, and second, the application of Neural Networks, which usually behave differently from the models we have used in this work.

We have compared our predictions to those performed by **2021arXiv210401875H**. They used data from 1.6 million SDSS DR12 galaxies to train seven different supervised models to predict redshift values. They have produced a set of metrics to assess their results, obtaining MAE~ 0.04, MSE~ 0.005, RMSE~ 0.071, R2~ 0.88, and Δz ~ 0.024.

When comparing with works which use AGN or QSO, we can take the results from

2021MNRAS.503.2639C, who compared the results of applying deep learning, decision trees, and k-nearest neighbours regression to predict redshift values for 100,000 SDSS DR12 QSO with accurate spectroscopic redshifts. For k-nearest neighbours, they quote RMSE~ 0.236 and $\Delta z \sim 0.034$, for decision tree regression, RMSE~ 0.333 and $\Delta z \sim 0.039$, and for deep learning, they obtained RMSE~ 0.235 and $\Delta z \sim 0.028$. Since they have used an SDSS sample, the properties of QSO among them are more homogeneous than that of the present work, leading to improved prediction results.

On a different approach, and to compare our results to a standard photometric redshift determination procedure, the redshift values predicted by our model can be contrasted with the values from Ananna et al. (2017). They used multi-wavelength data from 5,961 X-ray-detected AGN in the Stripe 82 Field with $z \leq 3.0$ and, from fitting SED models, they computed photometric redshifts. A value of $\Delta z \sim 0.041$ is reached for its full sample. And a catastrophic outlier fraction of 13.69% is achieved, less than half of what is obtained using our stacked model.

Model from pipeline

We compare our results to that of Ananna et al. (2017, Stripe 82X) where the authors analysed multi-wavelength data from more than 6 100 X-ray detected AGN from the 31.3 deg^2 of the Stripe 82X survey. They obtained photometric redshifts for almost 6 000 of these sources using the template-based fitting code `LePhare` (**1999MNRAS.310..540A**; **2006A&A...457..841I**). Their results present a normalised median absolute deviation of σ_{NMAD}=0.062 and an outlier fraction of η=13.69%, values which are similar to our results in HETDEX and S82 except for a better outlier fraction (as shown in Table 4.3, we obtain η_{S82} = 25.18%, $\sigma_{\mathrm{NMAD}}^{\mathrm{HETDEX}}$=0.071, and η^{HETDEX}=18.9%).

On the ML side, we compare our results to those produced by Carvajal et al. (2021) in S82, with $\sigma_{\mathrm{NMAD}} = 0.1197$ 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. **??**). 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 **2019PASP..131j8003L**. They studied the use of the k-nearest neighbours algorithm KNN (Cover and Hart, 1967), a non-parametric supervised learning approach, to derive redshift values for radio-detectable sources. They combined 1.4 GHz radio measurements, infrared, and optical photometry in the European Large Area ISO Survey-South 1 (ELAIS-S1; Oliver et al., 2000) and extended Chandra Deep Field South (eCDFS; Lehmer et al., 2005) fields, matching their sensitivities and depths to the expected values in the Evolutionary Map of the Universe (EMU; Norris et al., 2011). From the different experiments they run, their resulting NMAD values are in the range $\sigma_{\mathrm{NMAD}} = 0.05 - 0.06$, and their outlier fraction can be found between $\eta = 7.35\%$ and $\eta = 13.88\%$. As an extension to the previous results, **LUKEN2022100557** analysed multi-wavelength data from radio-detected sources the eCDFS and the ELAIS-S1 fields. Using KNN and RF methods to predict the redshifts of more than 1 300 RGs, they have developed regression methods that show NMAD values between $\sigma_{\mathrm{NMAD}} = 0.03$ and $\sigma_{\mathrm{NMAD}} = 0.06$, $\sigma_z = 0.10 - 0.19$, 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 $\sigma_{\mathrm{NMAD}} = 0.054$ and $\sigma_{\mathrm{NMAD}} = 0.48$ 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 HETDEX, it is possible to compare our results to those from **2019A&A...622A...3D**. They use a hybrid photometric redshift approach combining traditional template fitting redshift determination and ML-based methods. In particular, they implemented a Gaussian Process (GP) algorithm, which is able to model both the intrinsic noise and the uncertainties of the training features. Their redshift prediction analysis of AGN sources with a spectroscopic redshift detected in the LoTSS DR1 (6,811 sources) found a NMAD value of $\sigma_{\mathrm{NMAD}} = 0.102$ 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. **2019A&A...622A...3D** 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. 5.1.1). They combined three algorithms for their predictions: `XGBoost`, `CatBoost`, and `LightGBM` (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.

5.2 Influence of data imputation

One effect which might influence the training of the models and, consequently, the prediction for new sources is related to the imputation of missing values (cf. Sect. 2.1). In Fig. **??**, 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. **??** shows the influence of the degree of imputation in the classification between AGN and galaxies. For most of the bins, probabilities for predicted galaxies are distributed close to 0.0, without any noticeable trend. In the case of predicted AGN, the combination of low number of sources and high degree of imputation (`band_num` < 5) lead to low mean probabilities.

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

Finally, a stronger effect can be seen with the evolution of predicted redshift values for radio-detectable AGN. Despite the lower number of available sources, it is possible to recognise that sources with higher number of available measurements are predicted to have lower redshift values. Sources that are closer to us have higher probabilities to be detected in a large number of bands. Thus, it is expected that our model predicts lower redshift values for the most measured sources in the field.

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

5.3 Model explanations

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

5.3.1 Global feature importances

Overall, mean or global feature importances can be retrieved from models that are based on Decision Trees (**breiman2003manual**; e.g. Random Forests and Boosting models, Breiman, 2001). All algorithms selected in this work (`RF`, `CatBoost`, `XGBoost`, `ET`, `GBR`, and `GBC`) belong to these two classes. For each feature, the decrease in impurity (a term frequently used in the literature related to Machine Learning) of the dataset is calculated for all the nodes of the tree in which that feature is used. Features with the highest impurity decrease will be more important for the model (**NIPS2013_e3796ae8**)[1].

[1] 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. 3.2). Table 5.2 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 5.3) 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, i.e. algorithms used and stacking procedure. We first notice in Table 5.2 that the order of the features is different for all three models. This difference reinforces the need, as stated in Sect. **??**, 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 5.1.1). The WISE W3 photometry is though significantly less sensitive than W1, W2 or PS1 (see Fig. 2.5) and a significant number of sources will be represented as upper limits in such plot (see Table 2.2). From the importances in Table 5.2 and the values presented in Fig. 2.3 we infer that using optical colours could in principle create selection criteria with metrics equivalent to those shown in Table 5.1 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). We tested this hypothesis and derived a selection criterion in the g - r vs W1 - W2 colour-colour plot shown in Fig. **??** 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 5.1. Their limits are defined by the following expressions:

$$\mathbf{g-r} > \mathbf{-0.76}, \tag{5.1}$$

$$\mathbf{g-r} < \mathbf{1.8}, \tag{5.2}$$

$$\mathbf{W1-W2} > \mathbf{0.227\times(g-r)+0.43}, \tag{5.3}$$

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.

Table 5.2: 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		

Table 5.3: 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

One of the main potential uses of the pipeline is its capability to pinpoint radio-detectable AGN. The global features analysis for the radio detection model shows a high dependence on the near- and mid-IR magnitudes and colours, especially those coming from WISE. As a useful outcome similar to the AGN-Galaxy classification, we can use the most relevant features to build useful plots for the pre-selection of these sources and get insight into the origin of the radio emission. This is the case for the W4 histogram, shown in Fig. **??**, where sources predicted to be radio-emitting AGN extend to brighter measured W4 magnitudes. This added mid-IR flux might be simply due to an increased star formation rates (SFR) in these sources. In fact the $24\mu m$ flux is often used, together with that of $\mathrm{H}\alpha$ as a proxy for SFR (**2009ApJ...703.1672K**). The radio detection for these sources might have a strong component linked to the ongoing SF, especially for the sources with real or predicted redshift below $z{\sim}1.5$. A detailed exploration of the implications that these dependencies might have in our understanding of the triggering of radio emission on AGN, whether related to star formation (SF) or jets, is left for a future publication (Carvajal et al. in preparation).

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

Isolated redshift model

A first approach to understanding how the models process the data is calculating their feature importances. The use of features with high importance in a model will reduce the impurity of the classification or regression. That implies that a feature that can help the model to reduce its uncertainty over a prediction early on its training process will have a higher importance than a feature which can do it but later in the training.

We have obtained the feature importances for our three models and listed them in Table 5.4. It is also possible to see in these tables the final sets of features used by each model (see Section **??**).

For the stacked model, we can see that the features with the highest importances are those coming from the CatWISE catalogue. After them, quantities derived from Pan-STARRS

Table 5.4: Feature importances for the redshift prediction model.

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

observations. And, finally, those obtained from AllWISE, and GALEX observations, indicating a very low impact in the model training and the predictions derived from it.

Taking into account that CatWISE is the base catalogue from which all the sources in our sample have been drawn, it does not come as a surprise that features derived from it have the highest importance. None of their entries have been imputed, implying that they have the largest amount of relevant, non-repetitive information from all features, and consequently, the model tries to extract as much information as possible from the CatWISE data.

Despite the different nature of the used features –i.e., magnitudes, colours, ratios–, there is not a clear preference of one kind over the others. The main factor to have high importance is, as mentioned previously, the fraction of sources with a measurement in the studied feature.

5.3.2 Local feature importances: Shapley values

As opposed to the global (mean) assessment of feature importances derived from the decrease in impurity, local (i.e. source by source) information on the performance of such features can be obtained from Shapley values. This is a method from coalitional game theory that tells us how to fairly distribute the dividends (the prediction in our case) among the features (Shapley, 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 (**2020arXiv200805052M**). 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 were calculated using the python package `SHAP`[2] and, in particular, its module for Tree-based predictors (Lundberg et al., 2020). To speed calculations up, the package `FastTreeSHAP`[3] (**2021arXiv210909847Y**) was also used, which allows for multi-thread runs.

One way to display these SHAP values is through the so-called decision plots. They can show how individual predictions are driven by the inclusion of each feature. Besides determining the most relevant properties that help the model make a decision, it is possible to detect sources that follow different prediction paths which could be, eventually and upon further examination, labelled as outliers. An example of this decision plot, linked to the AGN-Galaxy classification, is shown in Fig. **??** for a subsample of the high-redshift ($z \geq 4.0$) spectroscopically classified AGN in the HETDEX field (121 sources, regardless of them being part of any sub-set involved in the training or validation of the models). The different features used by the meta-learner are stacked on the vertical axis with increasing weight and these final weight are sumarized in Table 5.5. Similarly, SHAP decision plots for the radio-detection and redshift prediction are presented in Figs. **??** and **??**, 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. Appendix **??** includes SHAP decision plots for all base learners used in this work. Additionally, and to be able to compare these results with the features importances from Sect. 5.3.1, we constructed Table 5.6, 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

[2] `https://github.com/slundberg/shap`

[3] `https://github.com/linkedin/fasttreeshap`

LoTSS). Table 5.6 shows, as Table 5.2, that the colour W1 - W2 is the most important discriminator between AGN and Galaxies for this specific set of sources. The importance of the rest of the features is mixed: similar colours are located on the top spots (e.g. g - r, W1 - W3 or r - i).

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

Isolated Redshift Model

As explained in Section 1, Shapley Values can be used to describe, for an individual member of a sample, the impact of the features on their predictions.

In our case, Shapley values were obtained using the Tree-based module of the Python package SHAP[4] (Lundberg and Lee, 2017; Lundberg et al., 2020), which is optimised for working with Tree-based models, as the ones used in this work.

In Figure **??**, the features are sorted by decreasing median Shapley values. As with feature importances, the quantity with the highest Shapley value is related to the base observations –CatWISE2020 data–. But from the distribution of values in the horizontal axis for the W1 magnitude, it is possible to see that its large dispersion implies that its influence on predictions can drive the final redshift either to low or high values. This is in contrast with, for instance, the $g - i$ colour. Its Shapley values might be close to zero or higher, indicating that, for individual sources, it does not have impact on the redshfit prediction or makes it have high redshift values.

The remaining features show a similar behaviour. Most of them show Shapley values clustered around 0.0, and a small sub-sample –generally, in a similar value range– deviates from this and have a noteworthy influence on predictions.

The feature with the second highest median Shapley values is the NUV magnitude from GALEX. From Figure **??**, it is possible to see that this feature exhibits a very high fraction

[4] https://github.com/slundberg/shap

of empty entries. That implies that most of sources have an imputed NUV magnitude. This distribution is present in Figure **??**, most of the sources are coloured with the strongest shade of red. Therefore, all imputed magnitudes make the redshift prediction go up, and all measured magnitudes make it go down.

Being able to retrieve these interpretations is one of the advantages of using Shapley values from a prediction model. It is possible to understand whether certain range of values of a feature can make a prediction go up or down.

Table 5.5: SHAP values (rescaled to add to 100) for base algorithms in each prediction step for observed features using 121 spectroscopically confirmed AGN at high redshift values ($z > 4$).

AGN-Galaxy model (`CatBoost`)			
Feature	SHAP value	Feature	SHAP value
`gbc`	36.250	`rf`	21.835
`et`	30.034	`xgboost`	7.198
	Remaining SHAP values:		4.683

Radio detection model (`GBC`)			
Feature	SHAP value	Feature	SHAP value
`rf`	11.423	`catboost`	5.696
`xgboost`	7.741	`et`	5.115
	Remaining SHAP values:		70.025

Redshift prediction model (`ET`)			
Feature	SHAP value	Feature	SHAP value
`xgboost`	41.191	`gbr`	13.106
`catboost`	20.297	`rf`	11.648
	Remaining SHAP values:		13.758

Table 5.6: 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		

This page intentionally left blank.

6

Conclusions

In this work, we trained several Machine Learning models to predict, from a sample of infrared-detected AGN –and their multi-wavelength counterparts– their redshift value. Apart from the photometric measurements, additional features were created using a set of photometric colours and flags to indicate whether a source exhibits or not radio or X-ray emission.

Sources were obtained from CatWISE2020 catalogue and counterpart measurements were obtained from AllWISE, Pan-STARRS, LOFAR, GMRT, VLASS, GALEX, 2MASS, and XMM-NEWTON observations and surveys.

Using of the `PyCaret` Python package as a framework, we stacked four different models with a meta-learner which took the outputs from them as extra features, harnessing the benefits of each individual model into one stronger prediction set.

The model was trained using the 90% of the AGN sample and applying it to the validation set lead a median redshift error on the prediction of $\Delta z = (z_{\mathrm{Predicted}} - z_{\mathrm{True}})/(1 + z_{\mathrm{True}}) = 0.0612$. This goes in line with previous results, taking into account that no major cleaning procedure was performed into the dataset.

Besides observations from WISE, which are the base catalogue for our sample –and thus, present in all sources–, the features that hold the largest importance to the model are those from Pan-STARRS.

In order to further test the predicting power of our model, we applied it to a separate catalogue of AGN located in the Stripe 82 Field and the median redshift error was $\Delta z = 0.0809$.

As a way to further understand the influence of the different features included in the model, Shapley values were calculated for all sources in the training sub-set. Their results match those from feature importance calculations, but also allow to see the impact of feature imputation. Most of imputed values have an impact that differs visibly from the non-imputed entries.

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

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

Our model can be further improved using future surveys which will cover large areas with very deep observations. One such survey is Data Release 2 (**LoTSS_DR2**) of the LoTSS survey. It will cover more than $5,600$ deg^2 in the northern sky with similar sensitivities as DR1. This will allow us training a similar redshift prediction model but with a number of sources one order of magnitude larger, improving its accuracy dramatically.

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

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

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

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

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

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

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

- Generalised stacking is a useful procedure which collects results from individual ML algorithms into a single model that can outperform each of the individual models, while preventing the inclusion of biases from individual algorithms. Proper selection of models and input features, together with detailed probability and threshold calibration maximises the metrics of the final model.

- Classification between AGN and galaxies derived from our model is in line with previous works. Our pipeline is able to retrieve a high fraction of previously-classified AGN from HETDEX (recall = 0.9621, precision = 0.9449) and from the S82 field (recall = 0.9401, precision = 0.9481).

- Radio detection classification for predicted AGN has proven to be highly demanding in terms of data needed for creating the models. Thanks to the use of the techniques shown in this article (i.e. feature creation and selection, generalised stacking, probability calibration, and threshold optimisation), we are able to retrieve previously-known radio-detectable AGN in the HETDEX field (recall = 0.5216, precision = 0.3528) and in the S82 field (recall = 0.5816, precision = 0.1229). These rates improve significantly upon a purely random selection (4 times better for the HETDEX field and 13 times better for S82), showing the power of ML methods for obtaining new RG candidates.

- The prediction of redshift values for sources classified to be radio-detectable AGN can deliver results that are in line with works that use either traditional or ML methods.

- Our models (classification and regression) can be applied to areas of the sky which have different radio coverage from that used for training without a strong degradation of the prediction results. This feature can lead to the use of our pipeline over very distinct datasets (in radio and multi-wavelength coverage) expecting to recover the sources predicted to be radio-detectable AGN with a high probability.

- Machine Learning models cannot be only used for a direct prediction of a value (or a set of 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 (**2020PASA...37...48M**; **2016mks..confE...6J**; e.g. RACS, EMU, and MIGHTEE; Norris et al., 2011, 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 (**2016mks..confE...1J**) or the future SKA where the population at the faintest sources will be dominated by star-forming galaxies. This change needs to use the corresponding data in the training set.

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

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

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. arXiv: 1801.03181 [astro-ph.IM].

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. arXiv: 1912.02905 [astro-ph.GA].

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. arXiv: 2207.01645 [astro-ph.IM].

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. eprint: https://www.tandfonline.com/doi/pdf/10.1080/00401706.1974.10489157. url: https://www.tandfonline.com/doi/abs/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.

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. arXiv: 1902.07982 [astro-ph.GA].

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. arXiv: 1710.01296 [astro-ph.GA].

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. arXiv: 2109.02713 [astro-ph.CO].

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. arXiv: 2209.13321 [astro-ph.GA].

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. arXiv: 1111.6619 [astro-ph.CO].

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. arXiv: 2005.03536 [astro-ph.HE].

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. arXiv: 1209.6055 [astro-ph.CO].

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. arXiv: 1706.09901 [astro-ph.GA].

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. arXiv: 0906.2173 [astro-ph.IM].

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. arXiv: 0804.3413 [astro-ph].

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.

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

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. arXiv: 1611.07526 [astro-ph.GA].

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. arXiv: 2107.02815 [astro-ph.GA].

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. arXiv: 2001.03135 [astro-ph.GA].

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. arXiv: 1812.06025 [astro-ph.HE].

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. arXiv: 1805.05222 [astro-ph.GA].

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. arXiv: 2010.08748 [astro-ph.GA].

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. arXiv: 2009.10727 [astro-ph.GA].

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. arXiv: 2101.01195 [astro-ph.GA].

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. arXiv: 1501.01982 [astro-ph.HE].

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

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. URL: https://www.frontiersin.org/article/10.3389/fspas.2021.658229.

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. URL: https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml.

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. arXiv: 1908.03720 [astro-ph.GA].

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. arXiv: 2105.11169 [astro-ph.IM].

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. arXiv: 2008.08099 [astro-ph.GA].

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

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. arXiv: 1912.02916 [astro-ph.GA].

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. arXiv: 2111.00778 [astro-ph.GA].

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. arXiv: 1612.05560 [astro-ph.IM].

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. arXiv: 1707.07690 [astro-ph.GA].

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. url: http://doi.acm.org/10.1145/2939672.2939785.

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., 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. arXiv: 2006.11287 [cs.LG].

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. arXiv: 2204.02080 [astro-ph.IM].

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. arXiv: 2203.03679 [astro-ph.CO].

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. arXiv: 2205.04587 [astro-ph.CO].

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. arXiv: 1308.5353 [astro-ph.CO].

Davies, L. J. M., Robotham, A. S. G., Driver, S. P. et al. (Oct. 2018). 'Deep Extragalactic VIsible Legacy Survey (DEVILS): motivation,design, and target catalogue'. In: MNRAS 480.1, pp. 768–799. doi: 10.1093/mnras/sty1553. arXiv: 1806.05808 [astro-ph.GA].

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. arXiv: 2012.05759 [astro-ph.GA].

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. url: https://doi.org/10.1038/s41598-021-92278-w.

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. arXiv: 2112.03939 [astro-ph.IM].

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. arXiv: 2110.01704 [astro-ph.GA].

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. arXiv: 1201.3899 [astro-ph.CO].

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

Dorogush, A. V., Gulin, A., Gusev, G. et al. (2017). 'Fighting biases with dynamic boosting'. In: *CoRR* abs/1706.09516. arXiv: 1706.09516. URL: http://arxiv.org/abs/1706.09516.

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. arXiv: 1009.0614 [astro-ph.CO].

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.

Euclid Collaboration, Bisigello, L., Conselice, C. J. et al. (June 2022). 'Euclid preparation: XXIII. Derivation of galaxy physical properties with deep machine learning using mock fluxes and H-band images'. In: *arXiv e-prints*, arXiv:2206.14944, arXiv:2206.14944. arXiv: 2206.14944 [astro-ph.GA].

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. arXiv: 2212.06907 [astro-ph.GA].

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. arXiv: 2105.12985 [astro-ph.GA].

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. arXiv: 1612.05243 [astro-ph.IM].

Frederiksen, T. F., Graur, O., Hjorth, J., Maoz, D. and Poznanski, D. (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. arXiv: 1211.2208 [astro-ph.CO].

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. URL: https://doi.org/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. URL: https://www.sciencedirect.com/science/article/pii/S0167947301000652.

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. arXiv: 1609.04153 [astro-ph.IM].

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. arXiv: 1305.1823 [astro-ph.CO].

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. arXiv: 1707.06052 [astro-ph.IM].

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. URL: https://doi.org/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. arXiv: 1812.07156 [astro-ph.IM].

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. URL: https://journals.ametsoc.org/view/journals/mwre/98/2/1520-0493_1970_098_0136_caotbp_2_3_co_2.xml.

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. arXiv: 2306.00068 [astro-ph.GA].

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.

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. arXiv: 1403.4620 [astro-ph.GA].

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. arXiv: 1501.01555 [astro-ph.GA].

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. arXiv: 2108.03271 [astro-ph.GA].

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. arXiv: 1806.04680 [astro-ph.GA].

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. arXiv: 1008.0658 [astro-ph.CO].

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. arXiv: 0806.0183 [astro-ph].

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., Richards, G. T. and Zeimann, G. R. (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. arXiv: 1103.5749 [astro-ph.CO].

Humphrey, A., Bisigello, L., Cunha, P. A. C. et al. (Sept. 2022). 'Euclid preparation: XXII. Selection of Quiescent Galaxies from Mock Photometry using Machine Learning'. In: *arXiv e-prints*, arXiv:2209.13074, arXiv:2209.13074. arXiv: 2209.13074 [astro-ph.IM].

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. arXiv: 2006.09476 [astro-ph.GA].

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. arXiv: 0809.2101 [astro-ph].

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. arXiv: 1911.05791 [astro-ph.GA].

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. arXiv: 1607.01190 [astro-ph.CO].

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. arXiv: 1405.7382 [astro-ph.GA].

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.

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.

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. arXiv: 1503.05206 [astro-ph.GA].

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. arXiv: 1711.03234 [astro-ph.GA].

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. url: https://doi.org/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.

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. arXiv: 2012.08857 [astro-ph.GA].

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. arXiv: 1907.01981 [astro-ph.IM].

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. arXiv: 1308.4190 [astro-ph.CO].

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. arXiv: astro-ph/0405604 [astro-ph].

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. arXiv: 2011.15030 [astro-ph.GA].

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.

Latimer, C. J., Reines, A. E., Hainline, K. N., Greene, J. E. and Stern, D. (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. arXiv: 2105.05876 [astro-ph.GA].

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. arXiv: 1403.3938 [astro-ph.CO].

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. arXiv: astro-ph/0506607 [astro-ph].

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. arXiv: 2110.13901 [astro-ph.GA].

Linardatos, P., Papastefanopoulos, V. and Kotsiantis, S. (2021). ‘Explainable AI: A Review of Machine Learning Interpretability Methods’. In: *Entropy* 23.1. ISSN: 1099-4300. DOI: 10.3390/e23010018. URL: https://www.mdpi.com/1099-4300/23/1/18.

Lisenfeld, U. and Völk, H. J. (Feb. 2000). ‘On the radio spectral index of galaxies’. In: A&A 354, pp. 423–430. arXiv: astro-ph/9912232 [astro-ph].

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. arXiv: 1506.08222 [astro-ph.GA].

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. arXiv: 2010.11202 [astro-ph.IM].

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. arXiv: 1905.03274 [astro-ph.IM].

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.

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. arXiv: 1812.07190 [astro-ph.GA].

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.

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. arXiv: 1406.6025 [astro-ph.EP].

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. arXiv: 1102.1996 [astro-ph.EP].

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. arXiv: 2012.13084 [astro-ph.IM].

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. arXiv: 1208.2530 [astro-ph.CO].

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.

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. arXiv: astro-ph/0607278 [astro-ph].

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. arXiv: 1511.07870 [astro-ph.GA].

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. arXiv: 2103.09246 [astro-ph.GA].

Messias, H., Afonso, J., Salvato, M., Mobasher, B. and Hopkins, A. M. (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. arXiv: 1205.4764 [astro-ph.CO].

Molnar, C. (2019). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book/..

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. arXiv: 2011.06001 [astro-ph.IM].

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. arXiv: 2208.10503 [astro-ph.CO].

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. arXiv: 2010.13857 [astro-ph.CO].

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. arXiv: 1309.5384 [astro-ph.CO].

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. arXiv: 1106.3219 [astro-ph.CO].

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. arXiv: 1902.05188 [astro-ph.IM].

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. arXiv: astro-ph/0003263 [astro-ph].

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. arXiv: 2212.01915 [astro-ph.GA].

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. arXiv: 1707.07134 [astro-ph.GA].

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. arXiv: 1609.00499 [astro-ph.GA].

— (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.

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.

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. arXiv: 1807.06209 [astro-ph.CO].

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

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. arXiv: 1412.6512 [astro-ph.IM].

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. arXiv: 2104.13333 [astro-ph.HE].

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. URL: https://doi.org/10.1057/jt.2009.5.

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. arXiv: 1811.05994 [astro-ph.IM].

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. arXiv: 1906.06974 [astro-ph.GA].

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. URL: https://doi.org/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. arXiv: 1811.05528 [astro-ph.GA].

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. arXiv: 2210.02307 [astro-ph.GA].

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. arXiv: 2008.03311 [astro-ph.IM].

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. arXiv: 1509.08483 [astro-ph.GA].

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. URL: https://doi.org/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. arXiv: 1811.07926 [astro-ph.GA].

Shy, S., Tak, H., Feigelson, E. D., Timlin, J. D. and Babu, G. J. (July 2022). 'Incorporating Measurement Error in Astronomical Object Classification'. In: AJ 164.1, 6, p. 6. DOI: 10.3847/1538-3881/ac6e64. arXiv: 2112.06831 [astro-ph.IM].

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. arXiv: 1006.4637 [astro-ph.CO].

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. arXiv: 1405.1737 [astro-ph.GA].

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.

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.

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. arXiv: 1205.0811 [astro-ph.CO].

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. arXiv: astro-ph/0410523 [astro-ph].

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. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1974.tb00994.x. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1974.tb00994.x.

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. arXiv: 2105.02434 [astro-ph.GA].

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. arXiv: 2010.11225 [astro-ph.GA].

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. arXiv: 1404.4937 [astro-ph.GA].

Tulio Ribeiro, M., Singh, S. and Guestrin, C. (Feb. 2016). '"Why Should I Trust You?": Explaining the Predictions of Any Classifier'. In: *arXiv e-prints*, arXiv:1602.04938, arXiv:1602.04938. doi: 10.48550/arXiv.1602.04938. arXiv: 1602.04938 [cs.LG].

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. arXiv: 2102.10706 [astro-ph.GA].

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. arXiv: 1305.3550 [astro-ph.IM].

van Rijsbergen, C. J. (1979). *Information Retrieval*. 2nd. USA: Butterworth-Heinemann. isbn: 0408709294.

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. url: https://doi.org/10.1007/978-3-030-05318-5_2.

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. arXiv: 2009.10721 [astro-ph.GA].

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. arXiv: 2010.00619 [astro-ph.CO].

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. arXiv: 2105.09171 [astro-ph.GA].

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. arXiv: astro-ph/0406223 [astro-ph].

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. arXiv: 1711.10504 [astro-ph.GA].

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. url: https://www.sciencedirect.com/science/article/pii/S0893608005800231.

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. arXiv: 1008.0031 [astro-ph.IM].

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. arXiv: 2304.06065 [astro-ph.GA].

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. eprint: https://academic.oup.com/biomet/article-pdf/87/4/954/633221/870954.pdf. URL: https://doi.org/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. arXiv: astro-ph/0006396 [astro-ph].

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.

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. arXiv: 1602.06294 [astro-ph.IM].

Appendices

A

Tables

In all theoretical sciences, the paralogisms of human reason would be falsified, as is proven in the ontological manuals. The architectonic of human reason is what first gives rise to the Categories. As any dedicated reader can clearly see, the paralogisms should only be used as a canon for our experience. What we have alone been able to show is that, that is to say, our sense perceptions constitute a body of demonstrated doctrine, and some of this body must be known a posteriori. Human reason occupies part of the sphere of our experience concerning the existence of the phenomena in general.

A.1 Name of Appendix Section

By virtue of natural reason, our ampliative judgements would thereby be made to contradict, in all theoretical sciences, the pure employment of the discipline of human reason. Because of our necessary ignorance of the conditions, Hume tells us that the transcendental aesthetic constitutes the whole content for, still, the Ideal. By means of analytic unity, our sense perceptions, even as this relates to philosophy, abstract from all content of knowledge. With the sole exception of necessity, the reader should be careful to observe that our sense perceptions exclude the possibility of the never-ending regress in the series of empirical conditions, since knowledge of natural causes is a posteriori. Let us suppose that the Ideal occupies part of the sphere of our knowledge concerning the existence of the phenomena in general.

A.2 Name of Second Appendix Section

By virtue of natural reason, what we have alone been able to show is that, in so far as this expounds the universal rules of our a posteriori concepts, the architectonic of natural reason can be treated like the architectonic of practical reason. Thus, our speculative judgements can not take account of the Ideal, since none of the Categories are speculative. With the sole exception of the Ideal, it is not at all certain that the transcendental objects in space and time prove the

Table A.1: Positions in League after 12 matches during Summer Season

Team	P	W	D	L	F	A	Pts
Manchester United	6	4	0	2	10	5	12
Celtic	6	3	0	3	8	9	9
Benfica	6	2	1	3	7	8	7
FC Copenhagen	6	2	1	3	5	8	7

validity of, for example, the noumena, as is shown in the writings of Aristotle. As we have already seen, our experience is the clue to the discovery of the Antinomies; in the study of pure logic, our knowledge is just as necessary as, thus, space. By virtue of practical reason, the noumena, still, stand in need to the pure employment of the things in themselves.

B

Individual Image

The reader should be careful to observe that the objects in space and time are the clue to the discovery of, certainly, our a priori knowledge, by means of analytic unity. Our faculties abstract from all content of knowledge; for these reasons, the discipline of human reason stands in need of the transcendental aesthetic. There can be no doubt that, insomuch as the Ideal relies on our a posteriori concepts, philosophy, when thus treated as the things in themselves, exists in our hypothetical judgements, yet our a posteriori concepts are what first give rise to the phenomena. Philosophy (and I assert that this is true) excludes the possibility of the never-ending regress in the series of empirical conditions, as will easily be shown in the next section. Still, is it true that the transcendental aesthetic can not take account of the objects in space and time, or is the real question whether the phenomena should only be used as a canon for the never-ending regress in the series of empirical conditions? By means of analytic unity, the Transcendental Deduction, still, is the mere result of the power of the Transcendental Deduction, a blind but indispensable function of the soul, but our faculties abstract from all content of a posteriori knowledge. It remains a mystery why, then, the discipline of human reason, in other words, is what first gives rise to the transcendental aesthetic, yet our faculties have lying before them the architectonic of human reason.

Figure B.1: Example image within the Appendix in LaTeX.