A multi-band AGN-SFG classifier for extragalactic radio surveys using machine learning

Karsten, J.; Wang, L.; Margalef-Bentabol, B.; Best, P.N.; Kondapally, R.; La Marca, A.; ... ; Sabater, J.

A&A 675, A159 (2023)
https://doi.org/10.1051/0004-6361/202346770

Astronomy & Astrophysics

A multi-band AGN-SFG classifier for extragalactic radio surveys using machine learning*

J. Karsten[1], L. Wang[1,2], B. Margalef-Bentabol[2], P. N. Best[3], R. Kondapally[3], A. La Marca[1,2], R. Morganti[1,4], H. J. A. Röttgering[5], M. Vaccari[6,7,8], and J. Sabater[3,9]

[1] Kapteyn Astronomical Institute, University of Groningen, Groningen 9747, AD, The Netherlands
e-mail: jesper.karsten1999@gmail.com; karsten@astro.rug.nl
[2] SRON Netherlands Institute for Space Research, Landleven 12, 9747 AD, Groningen, The Netherlands
[3] Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh, EH9 3HJ, UK
[4] ASTRON, the Netherlands Institute for Radio Astronomy, Oude Hoogeveensedijk 4, 7991 PD Dwingeloo, The Netherlands
[5] Leiden Observatory, Leiden University, PO Box 9513, 2300 RA Leiden, The Netherlands
[6] Inter-University Institute for Data Intensive Astronomy, Department of Astronomy, University of Cape Town, 7701 Rondebosch, Cape Town, South Africa
[7] Inter-University Institute for Data Intensive Astronomy, Department of Physics and Astronomy, University of the Western Cape, Robert Sobukwe Road, 7535 Bellville, Cape Town, South Africa
[8] INAF – Istituto di Radioastronomia, via Gobetti 101, 40129 Bologna, Italy
[9] UK Astronomy Technology Centre, Royal Observatory, Blackford Hill, Edinburgh, EH9 3HJ, UK

Received 28 April 2023 / Accepted 5 June 2023

ABSTRACT

Context. Extragalactic radio continuum surveys play an increasingly more important role in galaxy evolution and cosmology studies. While radio galaxies and radio quasars dominate at the bright end, star-forming galaxies (SFGs) and radio-quiet active galactic nuclei (AGNs) are more common at fainter flux densities.
Aims. Our aim is to develop a machine-learning classifier that can efficiently and reliably separate AGNs and SFGs in radio continuum surveys.
Methods. We performed a supervised classification of SFGs versus AGNs using the light gradient boosting machine (LGBM) on three LOFAR Deep Fields (Lockman Hole, Boötes, and ELAIS-N1), which benefit from a wide range of high-quality multi-wavelength data and classification labels derived from extensive spectral energy distribution (SED) analyses.
Results. Our trained model has a precision of 0.92±0.01 and a recall of 0.87±0.02 for SFGs. For AGNs, the model performs slightly worse, with a precision of 0.87±0.02 and a recall of 0.78±0.02. These results demonstrate that our trained model can successfully reproduce the classification labels derived from a detailed SED analysis. The model performance decreases towards higher redshifts, which is mainly due to smaller training sample sizes. To make the classifier more adaptable to other radio galaxy surveys, we also investigate how our classifier performs with a poorer multi-wavelength sampling of the SED. In particular, we find that the far-infrared and radio bands are of great importance. We also find that a higher signal-to-noise ratio in some photometric bands leads to a significant boost in the model performance. In addition to using the 150 MHz radio data, our model can also be used with 1.4 GHz radio data. Converting 1.4 GHz to 150 MHz radio data reduces the performance by ∼4% in precision and ∼3% in recall.

Key words. galaxies: active – methods: data analysis – catalogs

1. Introduction

Virtually all known massive galaxies host supermassive black holes (SMBHs) at their centres (Kormendy & Ho 2013). When such a black hole releases large amounts of energy by accreting gas rapidly, it can be observed as an active galactic nucleus (AGN). AGNs are of great importance in studying galaxy evolution because strong correlations exist between the SMBH mass and the physical properties of the host galaxy, such as its velocity dispersion and bulge mass (Ferrarese & Merritt 2000; Gebhardt et al. 2000; Kormendy & Ho 2013). In addition, the cosmic black hole accretion history is similar to the cosmic star formation history (Kormendy & Ho 2013). Theoretically, the energy released from AGNs can heat or expel the gas in the interstellar medium and quench the star formation activity in the host galaxies (a mechanism known as AGN feedback; Fabian 2012; King & Pounds 2015). This could explain why the galaxies we see today are not as bright or massive as we might expect them to be from models and numerical simulations, which do not include AGN feedback (Bower et al. 2006).

Radio continuum surveys play a critical role in the detection of AGNs, particularly in finding the jet-mode AGNs. Observations in the radio can detect synchrotron radiation powered by the central SBMHs and/or recent star formation activity. Early-type galaxies normally emit synchrotron radiation at $<4 \times 10^{20}$ W Hz^{-1} at GHz radio frequencies from interstellar relativistic electrons (Phillips et al. 1986; Sadler et al. 1989). Radio galaxies, on the other hand, have radio GHz emission at $>10^{22}$ W Hz^{-1} (Sadler et al. 1989) due to relativistic jets. In the past, only the bright end of the radio sky could be probed, resulting mostly in detections of radio-loud galaxies. However, with more sensitive surveys, the faint end of the radio sky can also be

* The final trained model is publicly available at https://github.com/Jesper-Karsten/MBASC

probed. This results in modern radio surveys being able to probe not just radio galaxies, but also radio-quiet AGNs (RQs) and star-forming galaxies (SFGs). Therefore, the need to efficiently and reliably classify different types of radio sources becomes increasingly more urgent.

Over the past few decades, many techniques have been developed to detect AGN activity in various parts of the electromagnetic spectrum. For example, the ratios of certain emission lines are different for some AGNs from the typical O-stars in non-radiative sources. This means that their ratios of line fluxes can be analysed to find AGNs using so-called Baldwin, Phillips & Terlevich (BPT) diagrams (Baldwin et al. 1981). In the mid-infrared (MIR), photometric information can be used to find dust emission from the obscuring molecular gas and dust surrounding the black hole, which peaks at a rest frame of a few microns (e.g. Stern et al. 2005; Donley et al. 2012), which divide sources into AGNs and SFGs. X-ray data can detect emission from the accretion disk corona, which indicates AGN activity. Radio continuum emission can be used to locate the jets of AGNs. Lastly, a spectral energy distribution (SED) analysis can be performed to detect an AGN, particularly if extensive multi-wavelength photometric information is available.

In terms of the classification scheme, AGNs can be classified into two categories based on their energetic output (Heckman & Best 2014). The first category includes AGNs whose energetic output is mostly released via electromagnetic radiation produced by radiatively efficient accretion of gas, which leads to the formation of an optically thick and geometrically thick accretion disk surrounding the SMBH (Shakura & Sunyaev 1973). This disk emits from the extreme ultraviolet (EUV) through the visible in the electromagnetic spectrum (Peterson 1997; Osterbrock & Ferland 2006; Krolik 1999). Additionally, this disk is surrounded by a hot corona, which Compton-up-scatters photons into the X-ray band. The ionising radiation from the disk and the corona heats and ionises a portion of the gas clouds surrounding the AGN. This results in the production of emission lines in the ultraviolet (UV), optical, and near-infrared (NIR). Lastly, the accretion disk is also surrounded by a cloud of molecular gas. A portion of UV, visible, and soft X-rays from the corona is absorbed by this dusty cloud and then emerges again as infrared emission. Traditionally, these AGNs are known as quasar-like AGNs. In this paper, we use the name 'high-excitation' (due to their strong high-excitation emission lines) or 'radiative-mode' AGNs.

The second category, known as jet-mode (or low-excitation) AGNs, consists of AGNs that produce less electromagnetic radiation compared to the first category. Their primary mode of energetic output is via kinetic energy transported in so-called jets (two-sided collimated beams of relativistic particles). It should be noted that a fraction of radiative-mode AGNs can also produce these jets. The geometrically thin accretion disk mentioned for the other type of AGN is either absent or is replaced by a geometrically thick structure (Quataert 2001; Ho 2008), which is consistent with the lower Eddington-scaled accretion rate. AGNs that have this excessive radio emission (as displayed by jets) are known as radio loud. They can be identified by their aforementioned jets or by observing an excess radio emission compared to what is expected based on star formation activity (Gürkan et al. 2018; Smith et al. 2021). The mechanism behind the generation of the jets is debated, but mechanisms involving rotating black holes and magnetic flux accretion are plausible (Condon & Mitchell 1984; Windhorst et al. 1985). At low flux densities (<0.1 mJy), the source counts are dominated by RQs and SFGs. At increasing flux densities, the source counts quickly become dominated by radio-loud AGN above ≈1 mJy (Padovani et al. 2015). These two binary criteria (radiative and radio excess) described above can then be used to define four classes: SFGs (non-radiative and no radio excess), RQs (radiative and no radio excess), low-excitation radio galaxy (LERG) (non-radiative and radio excess), and high-excitation radio galaxy (HERG) (radiative and radio-excess).

The main goal of this paper is to use supervised machine-learning (ML) trained on classification labels obtained from previous SED analyses to create a fast and reliable method of classifying radio sources as AGNs or SFGs. The advantage of ML algorithms is that once they are trained, it is quick and easy to apply them to a new similar dataset. In addition, ML classifications are always reproducible. We investigate supervised ML methods by using multi-wavelength photometry and photometric redshifts of radio sources detected in the first data release LOw-Frequency ARray (LOFAR) Two-metre Sky Survey (LoTSS) Deep Fields. The labels for these sources come from a detailed SED analysis with different SED fitting codes (Best et al. 2023).

This paper is organised as follows. In Sect. 2 we discuss the LOFAR radio data in the Deep Fields and the associated multi-wavelength photometric data on which the ML algorithm is trained. We also discuss how the separation between SFGs and AGNs was performed using an SED analysis. In Sect. 3 we describe the supervised ML algorithm adopted in this paper and the preprocessing, hyperparameters, and metrics we used. In Sect. 4 we present our results on the overall performance of the ML-based classifier, including a feature-relevance study. In addition, we investigate how the performance of the classifier depends on factors such as sample size, SED sampling, and signal-to-noise ratios (S/N) of the various filters. Finally, in Sect. 5, we present our conclusions of this study as well as information about the access to our classifier for radio sources.

2. Data

To apply supervised ML methods, labelled data are required. We used ∼80 000 radio sources in three LoTSS Deep Fields (Tasse et al. 2021; Sabater et al. 2021; Duncan et al. 2021; Kondapally et al. 2021): ELAIS-N1, Boötes, and Lockman Hole. These sources were cross-matched to their multi-wavelength counterparts. Best et al. (2023) performed an SED analysis using multiple fitting codes to classify the sources as SFG, RQ, HERG, or LERG. In this section, we present the key information regarding the LOFAR radio data and the associated multi-wavelength photometric data, as well as a brief summary of the SED-based classification process.

2.1. Parent radio source catalogues and the associated multi-wavelength data

Radio observations in the three fields were conducted using the LOFAR telescope (van Haarlem et al. 2013). This instrument performs deep and wide radio observations of the sky through its high-sensitivity high angular resolution and wide field of view. The LoTSS Deep Fields are a deep survey that includes the European Large Area Infrared Space Observatory Survey Northern Field 1 (ELAIS-N1; Oliver et al. 2000), the Boötes field (Jannuzi et al. 1999), and the Lockman Hole (Lockman et al. 1986). This survey has sufficient sky area to observe a full range of environments at wide redshift ranges, aiming to reach a noise level of 10–15 μJy beam^{-1} at 150 MHz. For the first data release, radio observations were taken with the High Band Antenna array (HBA) centred at roughly 150 MHz, and they are described by Tasse et al. (2021) for the Boötes and Lockman Hole fields and

by Sabater et al. (2021) for the ELAIS-N1 field. Source extraction is performed using the Python blob detector and source finder (Mohan & Rafferty 2015).

Each of these three fields has extensive associated multi-wavelength data across a wide range of the electromagnetic spectrum (Kondapally et al. 2021). We summarise the data available in each field here. For a detailed description of the multi-wavelength properties and cross identifications of the radio sources, we refer to Kondapally et al. (2021).

The far-ultraviolet (FUV) and near-ultraviolet (NUV) data come from data releases 6 and 7 of the Deep Imaging Survey (DIS) taken with the Galaxy Evolution Explorer (GALEX) space telescope (Martin et al. 2005; Morrissey et al. 2007) for all three fields. The GALEX observations cover around 13.5 deg^2 in ELAIS-N1, 8 deg^2 in Boötes, and also 8 deg^2 in Lockman Hole.

Observations in the *u* band are taken from the *Spitzer* Adaptation of the Red-sequence Cluster Survey (SpARCS; Wilson et al. 2009; Muzzin et al. 2009) in ELAIS-N1 and the Lockman Hole covering ∼12 and ∼13 deg^2. For Boötes, the *U*-band data were observed with the Large Binocular Telescope (LBT; Bian et al. 2013), which covers 9 deg^2.

In the optical, observations in the *grizy* bands were taken using the Panoramic Survey Telescope and Rapid Response System (PanSTARRS; Kaiser et al. 2010) in the Medium Deep Survey (MDS; Chambers et al. 2016) for ELAIS-N1. For Boötes, the *R* and *I* band are taken as part of the NOAO Deep Wide Field Survey (NDWFS; Jannuzi et al. 1999), and *z*-band data come from the zBoötes survey (Cool 2007), which covers the entire NDWFS field. Lastly, *y*-band data in Boötes were observed with the LBT covering the entire NDWFS field as well. The *g*-, *r*-, and *z*-band data were taken by SpARCs in the Lockman Hole, while *i* band was observed within the Red Cluster Sequence Lensing Survey (RCSLenS; Hildebrandt et al. 2016). MDS covers 8.05 deg^2 in ELAIS-N1, NDWFS covers 9.3 deg^2 in Boötes, and RCSLenS covers 16.63 deg^2 in Lockman Hole.

The NIR data in the *J* and *K* band come from the UK Infrared Deep Sky Survey Deep Extragalactic Survey (UKIDSS-DXS) Data Release 10 (Lawrence et al. 2007) for ELAIS-N1 (covering 8.87 deg^2) and the Lockman Hole (covering 8.16 deg^2). These observations were made using the WFCAM instrument (Casali et al. 2007) on the UK Infrared Telescope (UKIRT; Lawrence et al. 2007). For Boötes, the *J*-, *H*-, and *K*-band data were obtained within the NOAO Extremely Wide-Field Infrared Imager (NEWFIRM; Whitaker et al. 2011; Gonzalez 2010), covering 8.5 deg^2.

The MIR data at 3.6, 4.5, 5.8, and 8.0 µm come from the Infrared Array Camera (IRAC; Fazio et al. 2004) on the *Spitzer* Space Telescope (Werner et al. 2004) from the *Spitzer* Wide-area InfraRed Extragalactic (SWIRE) survey (Lonsdale et al. 2003) for ELAIS-N1 (covering 9.32 deg^2) and the Lockman Hole (covering 10.95 deg^2). On the same telescope, the *Spitzer* Deep Wide Field Survey (Ashby et al. 2009) observed filters from 3.6 to 8.0 µm for Boötes, covering approximately 10 deg^2.

The 24 µm data are taken using the Multi-band Imaging Photometer for *Spitzer* (MIPS; Rieke et al. 2004). They cover all fields.

Data at 100 and 160 µm were observed with the Photodetector Array Camera and Spectrometer (PACS; Griffin et al. 2010). 250, 350 and 500 µm were taken using the Spectral and Photometric Imaging Receiver (SPIRE; Poglitsch et al. 2010). All data were taken within the *Herschel* Multi-tiered Extragalactic Survey (HerMES; Oliver et al. 2012) by the *Herschel* Space Observatory (Pilbratt et al. 2010). They cover all three fields. These data are part of the *Herschel* Extragalactic Legacy Project (HELP; Shirley et al. 2021) with far-infrared (FIR) deblending for the radio sources described by McCheyne et al. (2022).

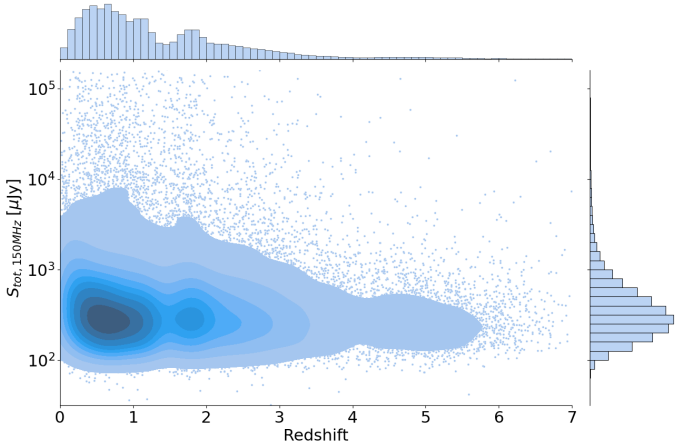

Fig. 1. Distribution of the 150 MHz radio flux vs photometric redshift. The histograms on the side give the distributions of the individual features as well.

The above paragraphs do not describe the full extent of the multi-wavelength data available in each field. We only include the data that we used. Some filters are only widely available in one field. We need consistent datasets over the field to use the ML algorithm on all three fields simultaneously, which gives us the maximum amount of data to train on. We therefore removed these filters. Similar filters are often available on different instruments (i.e. optical filters, e.g. *g*, *r*, *i*, *z*, and *y*). A choice was then made for the filter with the most complete data. This was done to limit the number of missing values because more complete data means a better performance of the model. Since not all fields have the same instruments and the same filters used to observe sources, some approximations had to be made. This means, in general, using similar filters or instruments to replace missing data (i.e. using the PanSTARRS *i*-band flux in ELAIS-N1 instead of NDWFS *I* band, which is used in Boötes). When no equivalent band was available, the feature was simply left empty. Non-detections and detections below 3σ were left empty. Table 1 shows the exact survey and corresponding depth for each field that was used for a specific feature.

For a minority of sources, spectroscopic redshifts are available (1602, 4039, and 1466 sources in ELAIS-N1, Boötes, and Lockman, respectively); for the other sources, photometric redshifts are necessary. Photometric redshifts in all fields were obtained by using a combination of template fitting and ML methods (Duncan et al. 2021). Duncan et al. (2021) used three template libraries: EAZY (Brammer et al. 2008), the Extended Atlas of Empirical SEDs (Brown et al. 2014), and the revised XMM-COSMOS team templates (Ananna et al. 2017). Additionally, they used the Gaussian process redshift code GPZ (Almosallam et al. 2016b,a). A final redshift was then obtained from these multiple different redshifts using a hierarchical Bayesian combination framework. The resulting redshifts have a very high accuracy, with a median scatter of $\Delta z/(1 + z_{\rm spec}) < 0.015$ for sources with $z < 1.5$.

To give a general impression of the wide dynamic range of data that are used in this paper, we plot the distribution of the radio 150 MHz flux densities versus redshift in Fig. 1. Data in other wavebands also extend over a wide range of redshifts and flux densities. A correlation matrix is plotted for the multi-wavelength photometric data (including the LOFAR radio

Table 1. Different filters and instruments used in each field.

ELAIS-N1 (~7.15 deg^2)	Boötes (~10.73 deg^2)	Lockman hole (~9.5 deg^2)
DIS FUV (26.3 [mag])	DIS FUV (26.3 [mag])	DIS FUV (26.3 [mag])
DIS NUV (26.7 [mag])	DIS NUV (26.7 [mag])	DIS NUV (26.7 [mag])
SpARCS *u* (25.4 [mag])	SpARCS *u* (25.9 [mag])	SpARCS *u* (25.5 [mag])
PanSTARRS *g* (25.5 [mag])	–	SpARCS g (25.8 [mag])
PanSTARRS *r* (25.2 [mag])	NDWFS *R* (25.2 [mag])	SpARCS *r* (25.1 [mag])
PanSTARRS *i* (25.0 [mag])	NDWFS *I* (24.6 [mag])	RCSLenS *i* (23.8 [mag])
PanSTARRS *z* (24.6 [mag])	zBoötes *z* (23.4 [mag])	SpARCS *z* (23.5 [mag])
PanSTARRS *y* (23.4 [mag])	LBT *y* (23.4 [mag])	–
UKIDSS-DXS *J* (23.2 [mag])	NEWFIRM *J* (23.1 [mag])	UKIDSS-DXS *J* (23.4 [mag])
–	NEWFIRM H (22.5 [mag])	–
UKIDSS-DXS *K* (22.7 [mag])	NEWFIRM *K* (20.2 [mag])	UKIDSS-DXS *K* (22.8 [mag])
SWIRE ch1 (23.4 [mag])	SDWFS ch1 (23.3 [mag])	SWIRE ch1 (23.4 [mag])
SWIRE ch2 (22.9 [mag])	SDWFS ch2 (23.1 [mag])	SWIRE ch2 (22.9 [mag])
SWIRE ch3 (21.2 [mag])	SDWFS ch3 (21.6 [mag])	SWIRE ch3 (21.2 [mag])
SWIRE ch4 (21.3 [mag])	SDWFS ch4 (21.6 [mag])	SWIRE ch4 (21.2 [mag])
MIPS24 (20 [μJy])	MIPS24 (20 [μJy])	MIPS24 (20 [μJy])
PACS100 (12.5 [mJy])	PACS100 (12.5 [mJy])	PACS100 (12.5 [mJy])
PACS160 (17.5 [mJy])	PACS160 (17.5 [mJy])	PACS160 (17.5 [mJy])
SPIRE250 (4 [mJy])	SPIRE250 (5 [mJy])	SPIRE250 (4 [mJy])
SPIRE350 (4 [mJy])	SPIRE350 (5 [mJy])	SPIRE350 (4 [mJy])
SPIRE500 (6 [mJy])	SPIRE500 (10 [mJy])	SPIRE500 (6 [mJy])
LoTSS (20 [μJy])	LoTSS (30 [μJy])	LoTSS (23 [μJy])

Notes. The 3σ depths in AB magnitudes are provided for the FUV to IRAC ch4 bands in brackets. These depths were estimated using variances from empty 3″ apertures. For the MIPS, PACS, and SPIRE bands, the limits at which fluxes can still be accurately deblended are given (McCheyne et al. 2022). For the radio data, the rms sensitivity is given.

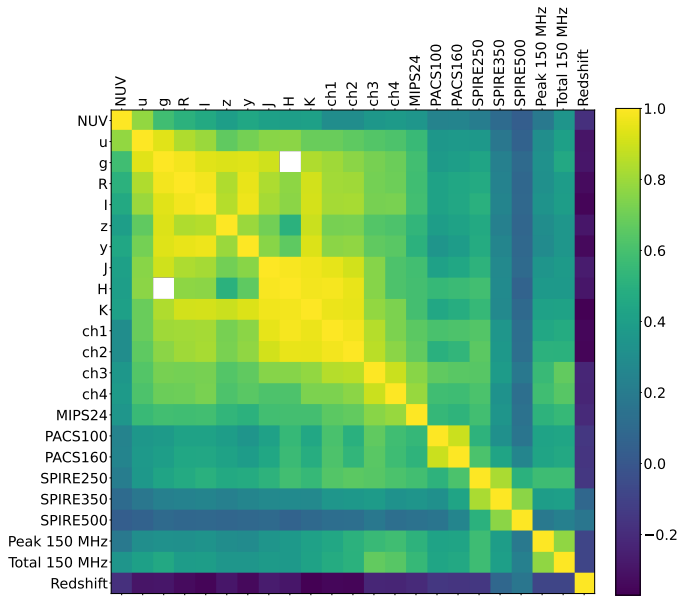

Fig. 2. Correlation matrix of all the features used as input for our ML classification. Only SFGs are included in this figure.

fluxes) in Fig. 2, which shows the linear correlation of the features. The figure has only been plotted for SFGs since adding AGNs would weaken the correlation between the infrared (IR) and the radio fluxes. This figure shows, as expected, that fluxes around similar wavelengths (i.e. between the NIR and the MIR) are more strongly correlated.

2.2. SFG-AGN classification

Using the photometric data and redshifts described in the previous section, Best et al. (2023) used four different SED fitting codes to classify sources as different classes of AGN or SFG. We briefly discuss each of the SED fitting codes and then discuss the final classification scheme. We refer to Best et al. (2023) for details.

The multi-wavelength analysis of galaxy physical properties (MAGPHYS; Da Cunha et al. 2008) and Bayesian analysis of galaxies for physical inference and parameter estimation (BAGPIPES; Carnall et al. 2018) codes are both SED fitting codes that assume energy balance. This means that the amount of energy absorbed at the optical and UV wavelengths by dust has to be the same as the energy emitted by the dust in the submillimeter and FIR. The main difference in the codes is their implementation of certain parametrisations and models. However, they generally give consistent results (Pacifici et al. 2023). Unfortunately, neither code includes AGN templates and therefore cannot provide reliable fits and parameters for galaxies in which the AGN contributes significantly to their UV to far-IR flux densities.

The code investigating galaxy emission (CIGALE; Boquien et al. 2019) is another model that uses an energy balance approach in SED fitting and modelling. It also includes AGN models, which makes the model significantly better for galaxies with significant AGN emission. The model incorporates the AGN light contribution, the IR emission from the heating of the dust by the AGN, and also the emission in the X-ray. Because of the additional parameters that follow from the AGN-fitting component, the model cannot sample the parameter space of the host galaxy properties as well as MAGPHYS and BAGPIPES for similar runtimes.

Finally, the version of AGNFITTER (Calistro Rivera et al. 2016) used by Best et al. (2023) does not use the principle of energy balance, but instead models four independent emission components. A blue bump, a stellar population, and an AGN torus with hot- and colder-dust emission. This way of fitting SEDs works better when the energy balance no longer holds (e.g. when the UV and FIR emissions are spatially offset from each other; Carnall et al. 2018). It can lead to aphysical solutions or poor constraints on the stellar population parameters, however.

Based on these various models, a set of selection criteria were applied by Best et al. (2023) to classify the sources as AGN or SFG and furthermore subdivide them into different AGN classes (HERG, LERG, and RQ). To classify a source as a radiative-mode AGN, two of three criteria have to be satisfied. First of all the 1σ lower limit of the AGN fraction (the fraction of IR luminosity from the contribution from the AGN dust torus component; referred to as P16) of the CIGALE fitting must be above 0.06 for ELAIS-N1 and Lockman Hole or 0.10 for Boötes. Secondly the P16 value from AGNFITTER must be above 0.16 for ELAIS-N1 and Lockman Hole or 0.25 for Boötes. Thirdly the lower reduced χ^2 value from MAGPHYS and BAGPIPES SED fits has to be greater than unity and a factor f greater than the lower reduced χ^2 of the CIGALE and AGNFITTER fits. This factor f was 1.36 for ELAIS-N1, 1.59 for the Lockman Hole, and 2.22 for Boötes.

The exact values of these cuts were derived by comparing the classifications to known secure classifications from spectroscopic and X-ray data and from classifications derived from MIR colour-colour diagrams. These criteria mean that a source is classified as a radiative-AGN if the AGN fraction is high in both CIGALE and AGNFITTER or if it only has a high AGN fraction in one of the SED fitting codes, but it has a very good SED fit.

In addition to classifying sources as a radiative-mode source, Best et al. (2023) also classified sources as radio loud or radio quiet using the radio data in the LOFAR Deep Fields. These radio-loud AGNs can be identified by analysing the correlation of SFGs between radio luminosity and their star-formation rate (SFR; Gürkan et al. 2018). Sources with a significantly higher radio luminosity than expected from this relation can then be classified as a radio-AGN. Best et al. (2023) used a ridgeline approach in which the sources are binned in narrow redshift bins, and within each bin, the mode of the distribution is picked as a ridgeline point. These ridgeline points can then be fitted with a linear relation. This results in the relation $\log(L_{150\mathrm{MHz}}) = 22.24 + 1.08\log(\mathrm{SFR})$, with $L_{150\mathrm{MHz}}$ in W Hz^{-1} and SFR in M_{sun} yr^{-1}. In ELAIS-N1 and the Lockman Hole, a source was deemed an AGN if it exceeded this ridgeline by 0.7 dex (about 3σ) and by $0.7+0.1z$ dex for Boötes. The relation is different in Boötes because in this field, the scatter increases at higher redshifts. A small percentage of sources cannot be classified using this method because the uncertainties at very low SFRs are large (below 0.01 $M_{\odot}$ yr^{-1}). Additionally, a few sources were not classified using this method (because they do not reach the radio excess threshold) but are clearly extended (>80 kpc) multi-component radio sources (incompatible with SFGs) from the LOFAR Galaxy Zoo project (Kondapally et al. 2021). These were added to the sample of radio-loud AGNs (about 0.5% of the total sample).

Based on the two subcriteria of radiative versus non-radiative and radio loud versus radio quiet, the four subclasses (SFG, RQ, HERG, or LERG) were derived. The results of this class division are listed in Table 2. This table shows that the data have a large imbalance within the classes: the sample contains 20 969 AGNs (27%) and 56640 SFGs (73%). For supervised ML methods, it can sometimes help to modify the dataset to reduce this imbalance. However, we opted not to do this because the performance did not improve. A brief discussion on this can be found in Appendix A.

Table 2. Class count in each field.

	SFG	LERG	RQ	HERG	AGN
ELAIS-N1	23 020 (76%)	4342 (14%)	2499 (8%)	387 (1%)	7228 (24%)
Boötes	12 213 (69%)	3219 (18%)	1906 (11%)	391 (2%)	5516 (31%)
Lockman Hole	21 407 (72%)	5206 (18%)	2465 (8%)	554 (2%)	8225 (28%)
Total	56 640 (73%)	12 767 (16%)	6870 (9%)	1332 (2%)	20 969 (27%)

3. Supervised ML classification of radio sources

Using the data and the labels described in the previous section, we trained a supervised ML algorithm on a two-class scheme (AGN or SFG). The aim of this model is to reproduce the labels using ML techniques.

3.1. Light gradient boosting machine

For the classification, we used the light gradient boosting machine (LightGBM or LGBM[1]; Ke et al. 2017). The LGBM uses a popular ML technique called gradient boosting. This ensemble technique uses multiple weaker learners (in the case of the LGBM, decision trees) to create a better model. Decision trees are structures in which a node at each depth poses a binary decision, for example, if the redshift is higher or lower than a given value. This leads to another pair of binary decisions, eventually ending in a classification. This results in 2^{n-1} nodes for a decision tree of depth n. Unlike random forests (Breiman 2001), which split up the dataset with a replacement to create multiple decision trees and then combine the results to predict the class, the LGBM works sequentially. Each weak learner (a decision tree) is fitted sequentially to reduce the error of the previous model. This loss can then be optimised sequentially using the gradient descent algorithm (Himmelblau 1972); hence the name gradient boosting. The loss we chose to optimise is the log loss, which is defined as

$$F = -\frac{1}{N}\sum_{i}^{N}\sum_{j}^{M} y_{ij}\cdot\log(p_{ij}). \quad (1)$$

Here N is the number of samples, M the number of different labels, y_{ij} is 1 if the instance belongs to the class and 0 if it does not, and p_{ij} is the probability of classifying instance i as label j. Gradient-boosted decision trees typically result in higher accuracies than a random forest (Li 2012). In contrast to other popular gradient boosting algorithms such as XGBoost (Chen & Guestrin 2016), the LGBM grows decision trees per leaf instead of per level (depth-wise). This difference ensures potentially higher accuracies, but can cause more overfitting.

[1] https://github.com/microsoft/LightGBM

Another advantage of the LGBM over random forests and similar techniques is that it can automatically deal with missing values. Since our data contain missing values for various features, this is an advantageous addition. The LGBM uses sparsity-aware split finding, which means that at each decision tree split, it assigns a missing value to the side that most reduces the loss. This allows the algorithm to obtain better accuracies on average on data with many missing values.

Since LGBM is a tree-based method, it is unnecessary to scale or normalise the redshifts and flux densities[2]. However, we have to create training, testing, and validation sets beforehand. This ensures that our model can classify the radio sources properly and does not only learn the structures on the data provided. To create these sets, we used the data we described before. We combined all of our three fields into one large dataset. This dataset was then split 80–20% to create a training and testing set. This testing set was split again 80–20% to create a validation set. The final proportions between training, testing, and validation sets were then 80 (62 087), 16 (9,934), and 4% (2483), respectively. The validation set was used for early stopping of the model. This means that the validation set was evaluated (but not trained on) at each training round of the model. If the performance on this validation set did not improve for ten rounds, the model was stopped. Since the model used the early stopping technique, tuning the number of rounds was unnecessary. Normally, the number of rounds needs to be tuned to ensure that the model does not train for too long, which reduces the performance. However, by using our validation set for early stopping, we can set the number of rounds (n_estimators in LGBM) to an arbitrarily high number (10^5).

3.2. Metrics

Proper metrics are necessary to accurately evaluate an ML model. The accuracy alone can give a false impression because it does not take the different datasets and the performance per class into account. Therefore, we additionally used metrics called the precision, recall, and F_1-score per dataset.

The precision, recall, and F_1-score are all metrics that range from 0 to 1, with 1 being the best (perfect classification) score, and 0 being the worst. They show the performance of the model per class instead of the overall performance, such as accuracy. The precision and recall are defined as (Olson & Delen 2008)

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{2}$$

and

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{3}$$

where TP are true positives, FP are false positives, and FN are false negatives. For AGNs, TP are the number of AGNs that are correctly classified as AGNs. FP are the number of classifications where SFGs are incorrectly classified as AGNs. FN are the number of classifications in which AGNs are incorrectly classified as SFGs. For SFGs, the inverse is true for TP, FP, and FN. The precision can then be described as the fraction of sources that are correctly classified as positive, while the recall can be described as the fraction of sources that were recalled. These two metrics can then be combined into the F_1-score as (Olson & Delen 2008)

$$F_1 = \frac{2}{\text{Recall}^{-1} + \text{Precision}^{-1}} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}}, \tag{4}$$

which is the harmonic mean of the precision and the recall. This gives a measure of the accuracy because actual accuracy is not possible class-wise.

[2] It is possible to add additional features to the model by combining different features into a new feature. Features such as colours could therefore be added. We opted not to do this because the colour information is also contained in the flux densities. The performance of the model does therefore not improve when colours are added.

Table 3. Hyperparameter search for LGBM.

	Search space	Optimal value
num_leaves	10–50	46
learning_rate	0.001–0.8	0.06369
min_data_in_leaf	1–20	1
colsample_bytree	0.1–1	0.5468
reg_alpha	0–5	2.619
reg_lambda	0–10	7.873

Notes. The search parameter space indicates the values in between which the optimal value is sought, and the optimal value is displayed on the right. To find the optimal value, we used a Bayesian optimisation algorithm. num_leaves is the maximum number of leaves a tree can have. learning_rate determines the step size during the learning process. min_data_in_leaf is the minimum number of samples in each decision leaf. colsample_bytree is the random subset of features the model trains on each iteration. reg_alpha and reg_lambda are L1 and L2 regularisation, respectively.

3.3. LGBM hyperparameters

The LGBM has a large number of hyperparameters that have to be optimised. These hyperparameters have a strong impact on the performance of the model. Instead of trying to find the best parameters by hand, we used Bayesian optimisation (Mockus 1975), using the BayesianOptimization python implementation[3]. This method tries to optimise a function by generating a posterior distribution. In our case, the function is the performance of the model based on the choice of hyperparameters. As more iterations are run, the posterior distribution improves. The method can then focus on exploring the regions in which it expects the output of the function (the model performance) to be highest. This allows for a much quicker and much more efficient search for the optimal hyperparameters.

The initial parameter space was taken to be a wide range of values, which is listed in Table 3. We then ran the Bayesian optimisation for 100 iterations. Each iteration cross-validated eight folds and took the average unweighted F_1-score as output. The highest unweighted F_1-score was chosen for the optimal hyperparameters, which are also listed in Table 3. For a detailed description of these parameters, the LGBM documentation can be consulted. A brief summary of each hyperparameter tuned can be found below Table 3.

3.4. Overfitting

Machine-learning models can overfit on the data used for training. Overfitting means that the ML algorithm learns structures on the training data too well and thus performs extremely well on that set, but it is not able to generalise to examples that were not used during the training. This can mean that the model learns

[3] https://github.com/bayesian-optimization/BayesianOptimization

Table 4. Results of the cross-validated two-class model on all the data and the individual fields.

All data			
	Precision	Recall	F1-score
SFG	0.92±0.01	0.96±0.01	0.94±0.01
AGN	0.87±0.02	0.79±0.02	0.83±0.02
Macro average	0.90±0.02	0.87±0.02	0.88±0.02
Weighted average	0.91±0.01	0.91±0.01	0.91±0.01

Boötes			
	Precision	Recall	F1-score
SFG	0.91±0.01	0.94±0.01	0.93±0.01
AGN	0.86±0.02	0.80±0.02	0.83±0.02
Macro average	0.89±0.02	0.87±0.02	0.88±0.02
Weighted average	0.90±0.01	0.90±0.01	0.90±0.01

Lockman Hole			
	Precision	Recall	F1-score
SFG	0.92±0.01	0.96±0.01	0.94±0.01
AGN	0.87±0.02	0.79±0.03	0.83±0.02
Macro average	0.90±0.02	0.87±0.03	0.88±0.02
Weighted average	0.91±0.02	0.91±0.01	0.91±0.01

ELAIS-N1			
	Precision	Recall	F1-score
SFG	0.93±0.01	0.96±0.01	0.94±0.01
AGN	0.87±0.02	0.77±0.03	0.82±0.03
Macro average	0.90±0.02	0.87±0.03	0.88±0.02
Weighted average	0.92±0.01	0.92±0.01	0.92±0.01

Notes. The macro average takes the unweighted mean of the two values above, resulting in a class-balanced metric. The weighted average weights each value by its fraction in the dataset. Values and errors are derived by the eight-fold cross-validation.

from the noise in the training data, for example. This results in very high performance metrics for the training set, but in a poor performance on the validation set.

Overfitting can be reduced by tuning the hyperparameters. As mentioned before, we also used a technique called early stopping, where the unweighted average F_1-score of the model was evaluated at each training round (epoch) on a set that is not seen during training. If the performance on the validation set does not improve for a certain number of epochs, the model stops training. Using this method, we can stop the model before it starts overfitting.

To investigate whether our model was overfitting, we considered the training histories of the model. Because the model runs iteratively, certain metrics perform in the training and validation set at each epoch. We used the log loss for this evaluation as this is also the loss that the model tries to minimise during training. When the training and validation set differ strongly, it can indicate overfitting of the model.

Figure 3 shows the log loss during the training process of the model. When the gap between the training and validation data becomes very wide, it is a strong indication of overfitting. The difference between the training and validation data is present but is not that large, it remains within a log loss of 0.1. Therefore, the training history does not indicate a significant amount of overfitting of our model.

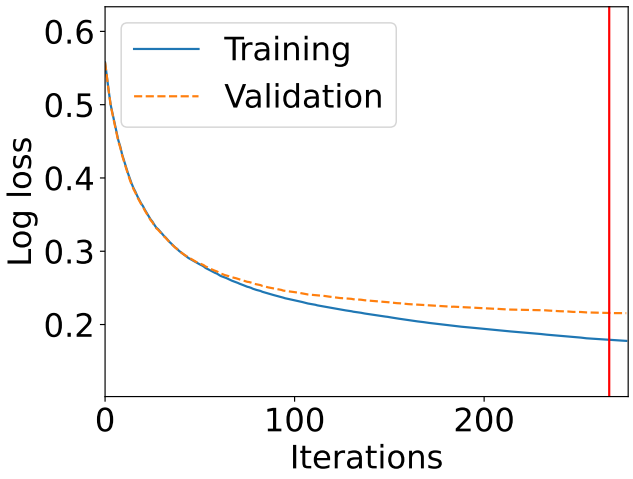

Fig. 3. Log loss for the validation and training set for each iteration during training. The training is stopped when the validation loss stopped to improve for ten iterations. This indicated by the vertical red line at 266 iterations.

4. Results

Using the hyperparameters described in Table 3, we trained our ML model. The model was cross-validated in an eight-fold stratified manner to keep the distribution between the three fields the same. In this section, we analyse how our trained model performs. To do this, averages and 1σ standard deviations were calculated and analysed.

4.1. Overall performance

Our model was trained on a binary classification scheme (AGN versus SFG). Best et al. (2023) provided four classes (HERG, LERG, RQ, or SFG) for their source classification, however, which means that the model can also be trained on four classes instead of two. We decided to focus on two classes in this paper because the performance on the four-class model is poorer. Even though the four-class model shows a similar accuracy as the two-class model described below, the performance on the minority classes is very poor. Particularly for the HERGs, the classifier reached very low (<50%) precision and recall. This poor performance is mostly due to the low number of sources in some classes. In Appendix C we summarise our investigations of a four-class model. In the main paper, we focus on the two-class model.

For the two-class model, Table 4 shows that our classifier has a total accuracy of 91%, which is a very good performance. This value is only representative of our class distribution (AGN or SFG); this value can be heavily biased if there is a strong class imbalance. In our case, our sample contained a large number of SFGs, which means that they affect the accuracy more than the AGNs. The larger fraction of SFGs affects the loss function of the model while training, and it thus results in a better performance for them compared to the AGNs. For the AGNs,

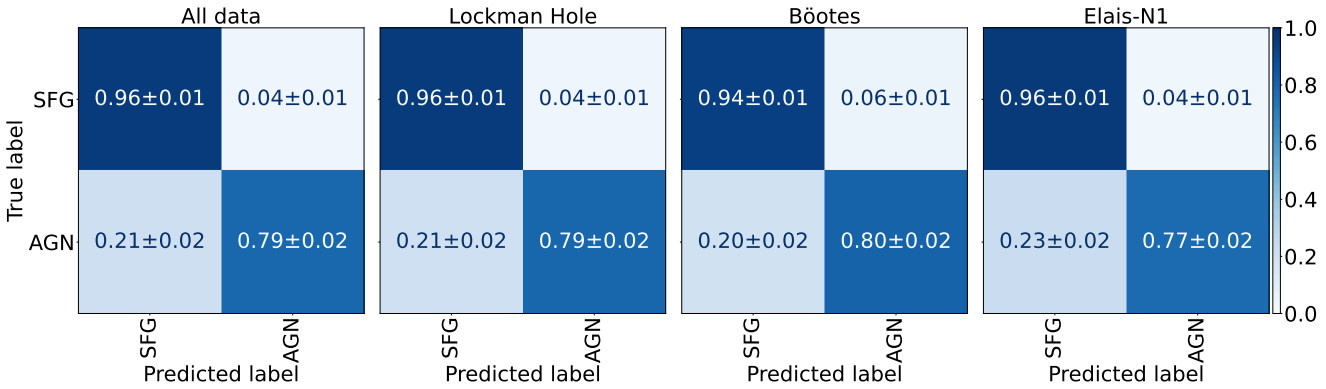

Fig. 4. Confusion matrices of the cross-validated two-class model. They show the cross-validated average fraction of how many SFGs and AGNS are classified correctly and incorrectly. A perfect classifier has all 1 across the diagonal and 0 everywhere else.

a precision of 87% is measured, which means that 87% of the sources classified as AGNs are true AGNs according to the labels used in this work. The recall is lower at 79%, meaning that we recover 79% of the AGNs in the data. The overall performance of the model can better be evaluated based on the unweighted (macro) averages of the metrics. The unweighted average is a good metric because it is not impacted by class imbalance. The macro F_1-score, which is a combination of both precision and recall, is 88%. Confusion matrices are plotted in Fig. 4. These show the fractions of how the model classifies sources. They were normalised over the rows, such that the diagonal represent the recalls. An ideal classifier has all 1s across the diagonal and 0s on the off-diagonal squares. Even for the minority class (the AGN), the performance holds up quite well, although about 22% of the AGNs are misclassified as SFGs.

The performance of the three classes of AGNs can also be analysed. This analysis was still made on the two-class model. We studied how well the subclasses were classified as AGN. The recall is 93%±3% for HERGs, 70%±3% for RQs, and 81%±2% for LERGs. Compared to the class distribution (12 767, 6870, and 1332 for LERGs, RQs, and HERGs, respectively), the HERGs seem to overperform as would be expected from their class size.

The analysis above is about the performance of our model using the three fields as training, validation, and testing sets. We are, however, also interested in the performance of the model on a new, unseen dataset. We simulated such a new dataset by only using two fields as training validation and testing sets and using another field purely as a testing set. We then compared the performances between the two testing sets to determine whether the model can generalise to new data. We used the Lockman Hole and Boötes data for our two fields and ELAIS-N1 as our testing field. The performance on the testing set when training on the two fields was approximately 2% lower in precision and recall and approximately 1% lower in accuracy. This indicates that our model would perform well on new data, provided the quality was similar to ELAIS-N1.

In addition to analysing the performance metrics, we also investigated the importance of the individual features. This not only helps identify the more important features, but also gives a better view of how all the features affect the model overall. Various methods exist for determining the feature relevance. They usually rely on some kind of score that each feature gives. We used shapley additive explanations[4] (SHAP) values (Lundberg & Lee 2017). SHAP gives each feature a value that describes its importance in the model. Additionally, it can show how features affect the model by studying the size and sign of the value. A higher and positive value means a higher impact, and a lower and negative value means a lower impact on the classification. Using the Python package created by Lundberg & Lee (2017), we determined the SHAP values for the model. In Fig. 5 we show the feature importance, where a higher SHAP value means that it is more likely to be an AGN. The figure mostly shows expected

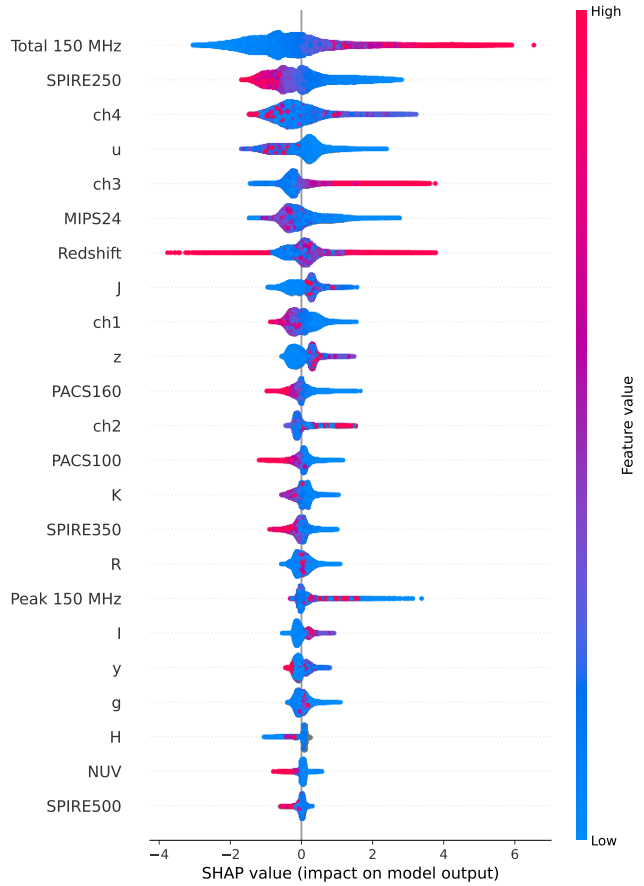

Fig. 5. Feature importance using SHAP values. The features are ordered by importance from top to bottom, with the most important feature being at the top. On the *x*-axis, the SHAP value is displayed. A positive value indicates a higher probability that the associated source is an AGN, while a negative value is a higher probability that the source is an SFG. The value of the feature is shown via the colour, which is also displayed on the right in a colour bar. For instance, a higher radio flux results in a higher probability that the source is an AGN.

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

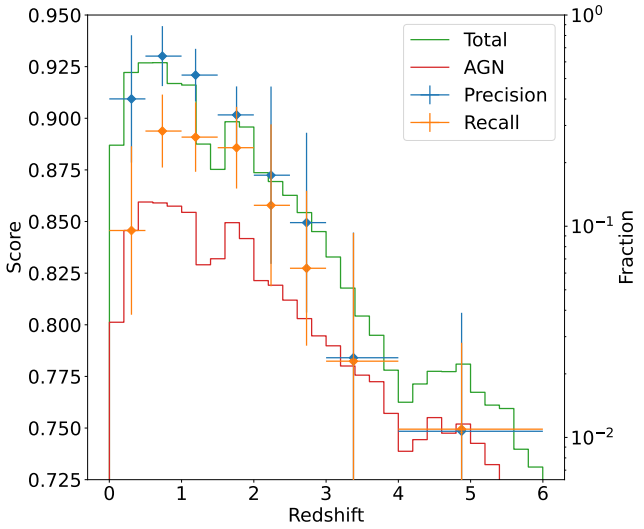

Fig. 6. Eight-fold cross-validated results of binned testing sets based on redshift. On the x-axis, we display the redshift. The points and errors are calculated by taking the mean and boundaries of each bin. In addition to the precision and recall on the left y-axis, the fraction of the data contained within the bin is plotted on the right y-axis. The y-errors represent 1σ standard deviations of the scores. The borders of the bins are [0, 0.5, 1, 1.5, 2, 2.5, 3, 4, 6].

results. Radio and IR features are generally most important, and fluxes in the visible are less important. Furthermore, higher radio fluxes mean that the source is more likely to be an AGN, which is expected for the radio-loud AGNs. This figure does not show any cross-interactions between the different features. It only displays how one feature affects the classifier overall.

4.2. Dependence on sample size, SED sampling, and S/N

The training data have more samples at lower redshift than at higher redshift. The model therefore learns the underlying structures at lower redshifts better because ML algorithms perform better with more data. Additionally, lower-redshift sources generally have higher-quality data than higher-redshift sources. This means that in addition to the sample-size dependence mentioned above, the metrics are worse due to the reduced data quality. SED classifications also become less reliable above $z = 2.5$ (Best et al. 2023). These three effects combined indicate that the performance degrades relatively quickly with increased redshift. To measure these effects, we performed eight-fold cross-validated tests in which we binned testing data by redshift and then measured the performance on them. These results are plotted in Fig. 6, where we also plot the corresponding fraction in a histogram. This plot clearly shows that the training size and score decrease when the redshift increases. The bin sizes were chosen manually such that each bin contained at least 5% of the training data.

Our model was trained on data with a certain number of features (fluxes and redshift). Other datasets, however, may not have the same features as those on which our model was trained on. It is therefore important to analyse how well our model performs when a testing set has fewer features.

The LGBM imputes automatic missing values. This technique is convenient when some values are missing, but it does not perform well when an entire feature (i.e. an entire column in the data) is missing. This is investigated in some detail in Appendix B. Therefore, when we wish to evaluate how well a model performs when a feature is missing, we cannot simply drop a column from the testing set and then evaluate the metric. Instead, we have to retrain the model without this column and evaluate the performance. We cannot give all possible combinations of missing features because if we have n features, the number of all possible combinations is 2^n (the power set) (Halmos 1960) for n features, which is extremely large in our case. Instead, we focus on some relatively common combinations and some combinations that affect the performance strongly.

Once again, we used an eight-fold cross-validation to calculate metrics. We removed the features from the training, validation, and testing set for each missing feature selection and then measured the performance. The results of this are listed in Table 5. The model performance does not degrade too much, except when we removed a large number of features. This means that we do not recommend using this model with very few features because then the performance is poor.

Lastly, the quality of the data can have a significant impact on the performance of the classifier. The quality of the data is measured by S/N. We calculated the S/N by dividing the fluxes by the errors provided with the multi-wavelength data. To measure the model performance for different S/N, we took binned S/N cuts in a particular wave band in the testing set and determined how performance differed for each bin. The model was trained on all the bins simultaneously to compare the different performances fairly for the bins.

To ensure that we can compare models fairly and objectively, we ensured that the main difference between each bin was the S/N and did not dependent on other factors. We therefore used adaptive bin sizes. This was done to ensure that each bin had the same number of sources in the training set. In general, this caused the lower S/N bins to become relatively narrow and the higher S/N bins to become wider because the sample size peaks at a relatively lower S/N. Each bin had a sample size of 5000. Because our total sample size was not a multiple of 5000, we discarded some very high S/N sources.

Because some of the flux densities are highly correlated, an increase in S/N in some bands increased the S/N of many bands simultaneously because the noise is largely uncorrelated. This means that the differences between S/N bins are significantly larger for these flux densities, while the performance difference might be minimum for other flux densities. The correlation between the different S/N of the features can be inferred from the correlation matrix in Fig. 2.

Using the abovementioned precautions, we ran the model on an eight-fold cross-validation and measured some of the macro-average precision and recall scores of the features. We chose a selection of features that showed limited linear correlation to investigate most of the spectrum. The results are plotted in Fig. 7. This figure shows that for certain bands such as the radio and IRAC channel 1, an increase in S/N results in a better performance of the model. For the g band, a positive trend is less significant, but still visible. For the IR features, an increase in S/N does not indicate an increase in performance, with even a possible decrease. This is in contrast to expectations, but could be due to uncertainties in the error estimates of these features.

4.3. Application to radio galaxies detected at 1.4 GHz

Because much research in radio astronomy is performed at 1.4 GHz, we also include a brief analysis of the performance of using the 1.4 GHz radio data instead of the 150 MHz radio

Table 5. Model performance when the model was retrained on fewer features.

	Precision SFG	Recall SFG	Precision AGN	Recall AGN	F_1-score
All	0.92 ± 0.01	0.96 ± 0.01	0.87 ± 0.02	0.78 ± 0.02	0.88 ± 0.02
NUV, *U*, *grizy*, *J*, *H*, *K*, ch1-ch4, MIPS, PACS, SPIRE	0.95 ± 0.01	0.89 ± 0.01	0.69 ± 0.01	0.83 ± 0.02	0.83 ± 0.01
NUV, *U*, *J*, *H*, *K*, ch1-ch4, MIPS, PACS, SPIRE	0.94 ± 0.01	0.89 ± 0.01	0.68 ± 0.01	0.82 ± 0.02	0.83 ± 0.01
NUV, *U*, *grizy*, *J*, *H*, *K*, ch3, ch4, MIPS, PACS, SPIRE	0.95 ± 0.01	0.89 ± 0.01	0.68 ± 0.01	0.82 ± 0.01	0.83 ± 0.01
NUV, *U*, *grizy*, *J*, *H*, *K*, MIPS, PACS, SPIRE	0.94 ± 0.01	0.87 ± 0.01	0.62 ± 0.01	0.81 ± 0.02	0.80 ± 0.02
NUV, *U*, *grizy*, *J*, *H*, *K*, PACS, SPIRE	0.94 ± 0.01	0.85 ± 0.01	0.56 ± 0.01	0.77 ± 0.01	0.77 ± 0.01
NUV, *U*, *grizy*, *J*, *H*, *K*, MIPS, SPIRE	0.95 ± 0.01	0.87 ± 0.01	0.61 ± 0.01	0.80 ± 0.02	0.80 ± 0.01
NUV, *U*, *grizy*, *J*, *H*, *K*, MIPS, PACS	0.94 ± 0.01	0.85 ± 0.01	0.55 ± 0.01	0.77 ± 0.01	0.77 ± 0.01
NUV, *U*, *grizy*, *J*, *H*, *K*	0.94 ± 0.01	0.81 ± 0.01	0.40 ± 0.01	0.72 ± 0.02	0.69 ± 0.02
grizy	0.94 ± 0.01	0.79 ± 0.01	0.34 ± 0.02	0.71 ± 0.02	0.66 ± 0.02
NUV, *grizy*, *J*, *H*, *K*, ch1-ch4, MIPS, PACS, SPIRE, 150 MHz	0.96 ± 0.01	0.92 ± 0.01	0.78 ± 0.01	0.87 ± 0.02	0.88 ± 0.01
NUV, *U*, *grizy*, ch1-ch4, MIPS, PACS, SPIRE, 150 MHz	0.96 ± 0.01	0.92 ± 0.01	0.78 ± 0.01	0.87 ± 0.02	0.88 ± 0.01
NUV, *U*, *J*, *H*, *K*, ch1-ch4, MIPS, PACS, SPIRE, 150 MHz	0.96 ± 0.01	0.92 ± 0.01	0.77 ± 0.01	0.86 ± 0.01	0.88 ± 0.01
NUV, *U*, *grizy*, *J*, *H*, *K*, MIPS, PACS, SPIRE, 150 MHz	0.95 ± 0.01	0.90 ± 0.01	0.71 ± 0.01	0.85 ± 0.02	0.85 ± 0.01
NUV, U, *grizy*, *J*, *H*, *K*, ch3, ch4, MIPS, PACS, SPIRE, 150 MHz	0.96 ± 0.01	0.92 ± 0.01	0.78 ± 0.01	0.86 ± 0.02	0.88 ± 0.01
NUV, *U*, *grizy*, *J*, *H*, *K*, ch1, ch2, MIPS, PACS, SPIRE, 150 MHz	0.96 ± 0.01	0.91 ± 0.01	0.74 ± 0.01	0.86 ± 0.02	0.86 ± 0.02
NUV, *U*, *grizy*, *J*, *H*, *K*, MIPS, PACS, SPIRE, 150 MHz	0.95 ± 0.01	0.90 ± 0.01	0.72 ± 0.01	0.85 ± 0.02	0.86 ± 0.02
NUV, *U*, *grizy*, *J*, *H*, *K*, 150 MHz	0.94 ± 0.01	0.84 ± 0.01	0.53 ± 0.01	0.77 ± 0.02	0.76 ± 0.02
NUV, *U*, *grizy*, *J*, *H*, *K*, ch1-ch4, PACS, SPIRE, 150 MHz	0.95 ± 0.01	0.91 ± 0.01	0.76 ± 0.01	0.85 ± 0.01	0.87 ± 0.02
NUV, *U*, *grizy*, *J*, *H*, *K*, ch1-ch4, MIPS, SPIRE, 150 MHz	0.96 ± 0.01	0.92 ± 0.01	0.78 ± 0.01	0.87 ± 0.02	0.88 ± 0.02
NUV, U, *grizy*, *J*, *H*, *K*, ch1-ch4, MIPS, PACS, 150 MHz	0.95 ± 0.01	0.91 ± 0.01	0.74 ± 0.01	0.84 ± 0.02	0.85 ± 0.01
NUV, *U*, *grizy*, *J*, *H*, *K*, ch1-ch4, 150 MHz	0.94 ± 0.01	0.88 ± 0.01	0.67 ± 0.01	0.82 ± 0.02	0.82 ± 0.02

Notes. The model was trained on an eight-fold cross-validation, and the precision, recall for each class and the total macro F_1-score were calculated. MIPS refers to MIPS24, PACS to PACS 100 and PACS160, and SPIRE to SPIRE250, SPIRE350, and SPIRE500.

data. We used the 1.4 GHz data available in the Lockman Hole from the Lockman Hole project (Prandoni et al. 2018). This set was cross-matched using a 3″ matching radius to the LOFAR Lockman Hole sample, resulting in 4005 matches. We compared the performance of the model using the LOFAR 150 MHz data versus predicted 150 MHz radio data from 1.4 GHz Lockman Hole project radio data. The predicted 150 MHz radio data were generated from the 1.4 GHz radio using a simple spectral index of $\alpha = 0.78$ derived by Mahony et al. (2016). These authors noted that this index becomes steeper with increasing flux densities (from $\alpha = 0.75$ to $\alpha = 0.84$). We did not change our spectral index with flux density because a detailed spectral index analysis is often not available on real data. The effect of this simpler approach is shown in Fig. 8, where the spectral index fits poorer at higher flux densities.

To ensure that our model did not train on the predicted 150 MHz fluxes, we simply used this new sample of 4005 sources as a testing set while training on the rest of the data. We compared the performance of this set with the true 150 MHz fluxes. The performance of the original 150 MHz sample reaches an accuracy of 91% and a macro-average F_1-score of 90%, similar to the expected values we know from Table 4. The performance of the model with the converted 1.4 GHz radio fluxes is listed in Table 6. The accuracy is 88%, and the macro-average F_1-score is 87%. Therefore, the performance is slightly worse, but it is still much better than simply dropping the radio features altogether (which results in a macro F_1-score of 83%).

Table 6. Model performance when it was trained with converted 1.4 GHz data.

	Precision	Recall	F1-score
SFG	0.89	0.90	0.89
AGN	0.86	0.85	0.85
Macro average	0.87	0.87	0.87
Weighted average	0.88	0.88	0.88

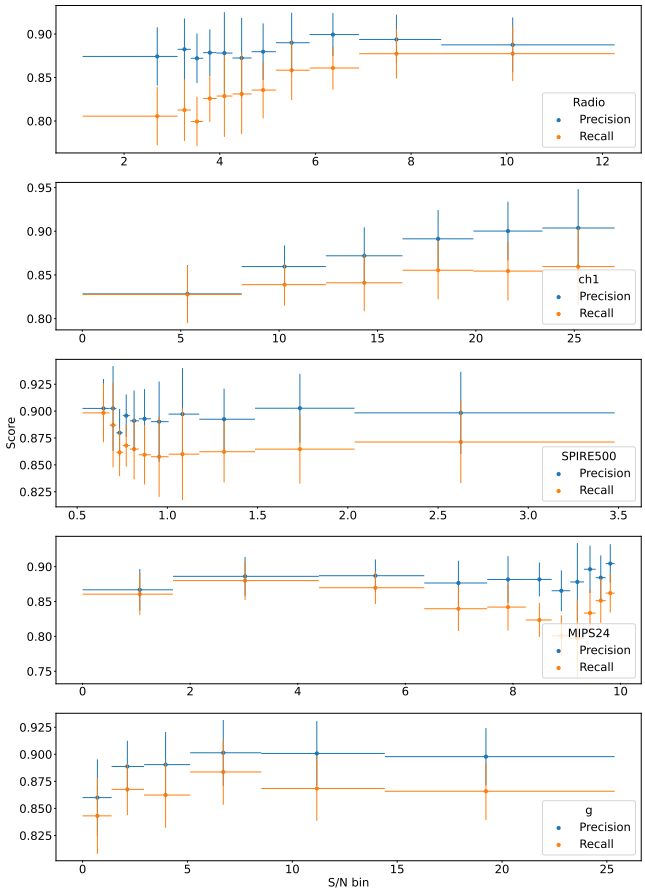

Fig. 7. Performance per S/N bin. The x-axis shows the mean value of the S/N bin. The y-axis denotes the macro average precision and recall. The uncertainties in the y-axis are calculated from the 1σ standard deviation over the eight-fold cross-validation. The uncertainty on the x-axis is the bin width.

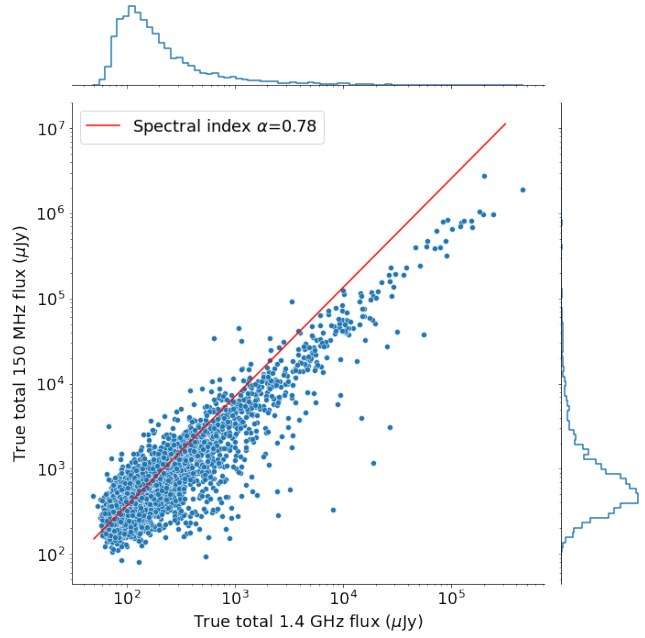

Fig. 8. Flux comparison of the observed LOFAR 150 MHz data and the observed 1.4 GHz data. The spectral index of 0.78 derived by Mahony et al. (2016) is plotted as the red line.

5. Conclusions

We created a supervised ML model to classify sources detected in extragalactic radio surveys as AGN or SFG. We used extensive radio and multi-wavelength data in three LoTSS Deep Fields: ELAIS-N1, Lockman Hole, and Boötes. Each field also had high-quality photometric redshifts. We combined these three fields by selecting features that are available in most of the fields, resulting in 77 609 sources, 20 969 of which are AGNs.

We created the ML classifier by using a decision-tree-based algorithm called LGBM. The eight-fold cross-validated testing resulted in an F_1-score of 0.94±0.01 for SFGs and 0.83±0.02 for AGNs, resulting in an average macro F_1-score of 0.88±0.06 and an accuracy of 0.91±0.01. We did not find significant deviations in the performance in each of the three different fields. The largest source of error in the model is that a fraction of 0.22±0.02 of the AGNs is misclassified as SFGs.

We tested the model performance for different sample sizes at different redshift bins. We find that lower sample sizes in the training set in redshift bins result in a reduced performance, which is expected. Furthermore, we investigated the model performance when fewer features were used. We retrained models without features and then tested the performance. We find that the performance decreases more when IR and radio features are removed, while removing the visible and UV features barely reduces the performance. Lastly, using S/N bins to see the model performance for different S/N values, we find that in general, a higher S/N results in some improvement of the model performance.

A public release of the model is available online[5]. This allows other researchers to use the model to classify AGNs in their own datasets.

References

Almosallam, I. A., Jarvis, M. J., & Roberts, S. J. 2016a, MNRAS, 462, 726
Almosallam, I. A., Lindsay, S. N., Jarvis, M. J., & Roberts, S. J. 2016b, MNRAS, 455, 2387
Ananna, T. T., Salvato, M., LaMassa, S., et al. 2017, ApJ, 850, 66
Ashby, M. L. N., Stern, D., Brodwin, M., et al. 2009, ApJ, 701, 428
Baldwin, J. A., Phillips, M. M., & Terlevich, R. 1981, PASP, 93, 5
Best, P. N., Kondapally, R., Williams, W. L., et al. 2023, MNRAS, 523, 1729
Bian, F., Fan, X., Jiang, L., et al. 2013, ApJ, 774, 28
Boquien, M., Burgarella, D., Roehlly, Y., et al. 2019, A&A, 622, A103
Bower, R. G., Benson, A. J., Malbon, R., et al. 2006, MNRAS, 370, 645
Brammer, G. B., van Dokkum, P. G., & Coppi, P. 2008, ApJ, 686, 1503
Breiman, L. 2001, Mach. Learn., 45, 5
Brown, M. J. I., Moustakas, J., Smith, J. D. T., et al. 2014, ApJS, 212, 18
Calistro Rivera, G., Lusso, E., Hennawi, J. F., & Hogg, D. W. 2016, ApJ, 833, 98
Carnall, A. C., McLure, R. J., Dunlop, J. S., & Davé, R. 2018, MNRAS, 480, 4379
Casali, M., Adamson, A., Alves de Oliveira, C., et al. 2007, A&A, 467, 777
Chambers, K. C., Magnier, E. A., Metcalfe, N., et al. 2016, ArXiv e-prints [arXiv:1612.05560]
Chen, T., & Guestrin, C. 2016, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16 (New York, NY, USA: ACM), 785
Condon, J. J., & Mitchell, K. J. 1984, AJ, 89, 610
Cool, R. J. 2007, ApJS, 169, 21
Da Cunha, E., Charlot, S., & Elbaz, D. 2008, MNRAS, 388, 1595
Donley, J. L., Koekemoer, A. M., Brusa, M., et al. 2012, ApJ, 748, 142
Duncan, K. J., Kondapally, R., Brown, M. J. I., et al. 2021, A&A, 648, A4
Fabian, A. C. 2012, ARA&A, 50, 455
Fazio, G. G., Hora, J. L., Allen, L. E., et al. 2004, ApJS, 154, 10
Ferrarese, L., & Merritt, D. 2000, ApJ, 539, L9
Gebhardt, K., Bender, R., Bower, G., et al. 2000, ApJ, 539, L13
Gonzalez, A. 2010, AAS Meeting Abstracts, 216, 415.13

[5] https://github.com/Jesper-Karsten/MBASC

Griffin, M. J., Abergel, A., Abreu, A., et al. 2010, A&A, 518, L3
Gürkan, G., Hardcastle, M. J., Smith, D. J. B., et al. 2018, MNRAS, 475, 3010
Halmos, P. R. 1960, Naive Set Theory (Princeton: The University Series in Undergraduate Mathematics), 104
Heckman, T. M., & Best, P. N. 2014, ARA&A, 52, 589
Hildebrandt, H., Choi, A., Heymans, C., et al. 2016, MNRAS, 463, 635
Himmelblau, D. M. 1972, Appl. Nonlinear Programming (New York: McGraw-Hill)
Ho, L. C. 2008, ARA&A, 46, 475
Jannuzi, B. T., Dey, A., & NDWFS Team. 1999, in AAS Meeting Abstracts, 195, 12.07
Kaiser, N., Burgett, W., Chambers, K., et al. 2010, SPIE Conf. Ser., 7733, 77330E
Ke, G., Meng, Q., Finley, T., et al. 2017, in Advances in Neural Information Processing Systems, eds. I. Guyon, U. V. Luxburg, S. Bengio, et al. (USA: Curran Associates, Inc.), 30
King, A., & Pounds, K. 2015, ARA&A, 53, 115
Kondapally, R., Best, P. N., Hardcastle, M. J., et al. 2021, A&A, 648, A3
Kormendy, J., & Ho, L. C. 2013, ARA&A, 51, 511
Krolik, J. H. 1999, Active Galactic Nuclei: From the Central Black Hole to the Galactic Environment (Princeton: Princeton University Press)
Lawrence, A., Warren, S. J., Almaini, O., et al. 2007, MNRAS, 379, 1599
Li, P. 2012, ArXiv e-prints [arXiv:1203.3491]
Lockman, F. J., Jahoda, K., & McCammon, D. 1986, ApJ, 302, 432
Lonsdale, C. J., Smith, H. E., Rowan-Robinson, M., et al. 2003, PASP, 115, 897
Lundberg, S. M., & Lee, S.-I. 2017, in Advances in Neural Information Processing Systems 30, eds. I. Guyon, U. V. Luxburg, S. Bengio, et al. (USA: Curran Associates, Inc.), 4765
Mahony, E. K., Morganti, R., Prandoni, I., et al. 2016, MNRAS, 463, 2997
Martin, D. C., Fanson, J., Schiminovich, D., et al. 2005, ApJ, 619, L1
McCheyne, I., Oliver, S., Sargent, M., et al. 2022, A&A, 662, A100
Mockus, J. 1975, IFAC Proc. Vol., 8, 428
Mohan, N., & Rafferty, D. 2015, Astrophysics Source Code Library [record ascl:1502.007]
Morrissey, P., Conrow, T., Barlow, T. A., et al. 2007, ApJS, 173, 682
Muzzin, A., Wilson, G., Yee, H. K. C., et al. 2009, ApJ, 698, 1934
Oliver, S., Rowan-Robinson, M., Alexander, D. M., et al. 2000, MNRAS, 316, 749
Oliver, S. J., Bock, J., Altieri, B., et al. 2012, MNRAS, 424, 1614
Olson, D., & Delen, D. 2008, Advanced Data Mining Techniques (Berlin: Springer)
Osterbrock, D. E., & Ferland, G. J. 2006, Astrophysics of Gaseous Nebulae and Active Galactic Nuclei (USA: AIP)
Pacifici, C., Iyer, K. G., Mobasher, B., et al. 2023, ApJ, 944, 141
Padovani, P., Bonzini, M., Kellermann, K. I., et al. 2015, MNRAS, 452, 1263
Peterson, B. M. 1997, An Introduction to Active Galactic Nuclei (Cambridge: Cambridge University Press)
Phillips, M. M., Jenkins, C. R., Dopita, M. A., Sadler, E. M., & Binette, L. 1986, AJ, 91, 1062
Pilbratt, G. L., Riedinger, J. R., Passvogel, T., et al. 2010, A&A, 518, L1
Poglitsch, A., Waelkens, C., Geis, N., et al. 2010, A&A, 518, L2
Prandoni, I., Guglielmino, G., Morganti, R., et al. 2018, MNRAS, 481, 4548
Quataert, E. 2001, ASP Conf. Ser., 224, 71
Rieke, G. H., Young, E. T., Engelbracht, C. W., et al. 2004, ApJS, 154, 25
Sabater, J., Best, P. N., Tasse, C., et al. 2021, A&A, 648, A2
Sadler, E. M., Jenkins, C. R., & Kotanyi, C. G. 1989, MNRAS, 240, 591
Shakura, N. I., & Sunyaev, R. A. 1973, A&A, 24, 337
Shirley, R., Duncan, K., Campos Varillas, M. C., et al. 2021, MNRAS, 507, 129
Smith, D. J. B., Haskell, P., Gürkan, G., et al. 2021, A&A, 648, A6
Stern, D., Eisenhardt, P., Gorjian, V., et al. 2005, ApJ, 631, 163
Tasse, C., Shimwell, T., Hardcastle, M. J., et al. 2021, A&A, 648, A1
van Haarlem, M. P., Wise, M. W., Gunst, A. W., et al. 2013, A&A, 556, A2
Werner, M. W., Roellig, T. L., Low, F. J., et al. 2004, ApJS, 154, 1
Whitaker, K. E., Labbé, I., van Dokkum, P. G., et al. 2011, ApJ, 735, 86
Wilson, G., Muzzin, A., Yee, H. K. C., et al. 2009, ApJ, 698, 1943
Windhorst, R. A., Miley, G. K., Owen, F. N., Kron, R. G., & Koo, D. C. 1985, ApJ, 289, 494

Appendix A: Data imbalance

Table 2 shows that the class sizes differ. The SFGs account for about 66% of the complete dataset. Imbalanced datasets can impact the performance of the model. This is because the model then learns more from the majority class but does not learn much from the minority classes. Additionally, it makes analyses of the model harder because most performance metrics are mostly influenced by the majority class.

A simple option to remedy this is to assign class weights to sources based on their class, such that a class that is x times more frequent than some other class has a weight of 1/x. This weight then affects the score that the algorithm calculates for its gradient descent. The LGBM provides a simple sample_weight argument for this. Unfortunately, when this parameter is used in the Bayesian optimisation, it reduces the accuracy and F_1-score by about 1%. This is a common occurrence because sometimes adding an additional sample weight can cause the model to learn less well on the larger classes.

Another relatively simple approach is to remove sources such that the data are more balanced. This option is not viable in our case because removing sources harms the model performance more (because there are fewer data to train on) than it improves due to a more balanced class distribution.

Lastly, we also tried to generate additional data. We did this by using a relatively simple approach where we generated additional sources using the data we already had. We generated new sources by using normal distributions with the errors on each flux and the redshift as the standard deviation. Unfortunately, when we generated new data such that all classes were balanced, the model performance did not improve. The lack of improvement can be explained by the fact that these Gaussian-generated sources are still quite similar to the original sources. They do not convey new complex information about what these sources could be. This approach might actually increase the degree of overfitting on the data because it essentially adds noisy copies of the data to the training set.

Appendix B: Missing values

Even though the LGBM allows the automatic handling of missing values, this feature does not work well with the entire missing features (columns) in the data. This fact can be shown qualitatively by manually removing values in certain columns, measuring the performance on the same model, and comparing this against a retrained model in which the same columns were dropped in the training and validation set. The results are shown in Table B.1. These results are not cross-validated because retraining the model each time is very time-consuming. Still, they show qualitatively that most features improve qualitatively when the model is in general retrained. For this reason, we employed a strategy of retraining the model when we tested the performance for missing features. This approach is not always necessary, however, because the improvement for some features is minimal.

Appendix C: Four-class model

In addition to training a two-class model, we also trained a four-class model. This model was trained in a similar way to the two-class model. As for the two-class model, we plot a confusion matrix to give an overview of the model. This confusion matrix is plotted in Fig. C.1. This figure shows that the boundary between SFG and AGN is just as good as in the two-class model. However, the classification of the subclasses of AGN performs far worse. For the HERG class, a recall of only 38%±6% is achieved. For the other classes, higher scores are displayed, but are still insufficient for an accurate classifier.

A large fraction of the misclassifications are AGN subclasses that are misclassified as SFG. This is because SFGs are the majority class. Additionally, the HERGs are often misclassified only in radio-loud mode (i.e. HERGs are classified as LERGs) or radiative mode (i.e. HERGs are classified as RQs).

The poor performance of the minority classes is partly due to the small size of these classes. This class imbalance could be fixed by removing a large portion of the larger classes, but this would result in very few remaining sources. The training set would then become too small to achieve a good performance. Additional data, particularly for the minority classes, are thus required.

Table B.1. Comparison of the macro F_1-score performance of the model for different missing-feature strategies.

	Only dropped	Retrained
No missing	0.88	-
LOFAR 144 MHz	0.77	0.84
u	0.86	0.88
J, H, K	0.87	0.88
g, r, i, z, y	0.86	0.88
ch1, ch2	0.83	0.88
ch3, ch4	0.83	0.87
ch1, ch2, ch3, ch4	0.74	0.86
MIPS24	0.85	0.86
PACS100, PACS160	0.87	0.87
SPIRE250, SPIRE350, SPIRE500	0.83	0.85

Notes. The left column shows the model performance that was not retrained after one or multiple features were removed, but the right column was retrained.

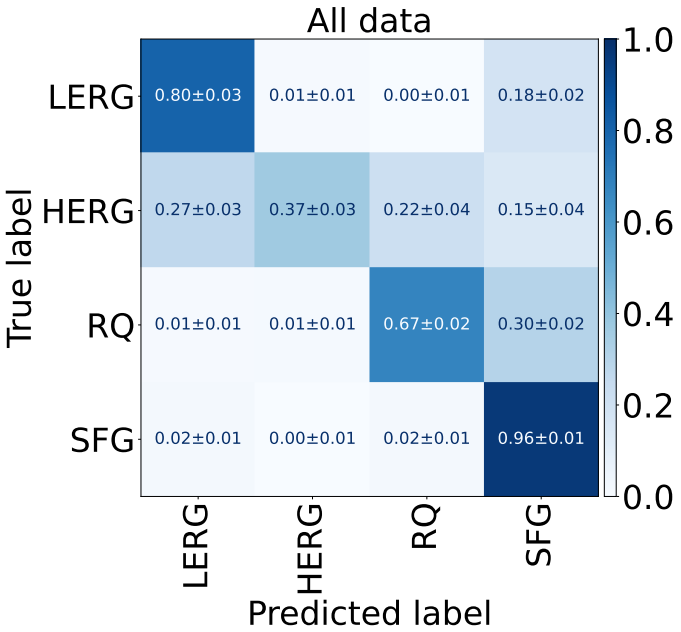

Fig. C.1. Confusion matrix of the cross-validated four-class model. This confusion matrix has been created on the cross-validated testing sets. A perfect classifier has all 1s across the diagonal and 0s everywhere else. It has been normalised over the rows, such that the diagonal represents the recalls.