Publications of the Astronomical Society of Australia (2026), **43**, e047, 17 pages
doi:10.1017/pasa.2026.10175

CAMBRIDGE UNIVERSITY PRESS

Research Article

EMU/GAMA: A statistical perspective on active galactic nuclei diagnostics

Jahang Prathap[1,2,3], Andrew Hopkins[1,2], Rodrigo Carvajal[4,5], Michael Cowley[6,7], Scott Croom[8], Duncan Farrah[9,10], Isabella Prandoni[11], Stanislav Shabala[12], Jacco Th. van Loon[13], Ciro Pappalardo[4,5], Kevin Pimbblet[14,15], Ummee Tania Ahmed[2,16,17], Maciej Bilicki[18], Michael Brown[19], Denis Leahy[20], Anilkumar Mailvaganam[1,2], Joshua Marvil[21], Tamal Mukherjee[1,2], Syed Faisal ur Rahman[22,23], Tessa Vernstrom[3,24], Jayde Willingham[1,2] and Tayyaba Zafar[1,2]

[1]School of Mathematical and Physical Sciences, Macquarie University, Sydney, NSW, Australia, [2]Astrophysics and Space Technologies Research Centre, Macquarie University, Sydney, NSW, Australia, [3]Australia Telescope National Facility, CSIRO, Space and Astronomy, Bentley, WA, Australia, [4]Instituto de Astrofísica e Ciências do Espaço, Universidade de Lisboa, OAL, Tapada da Ajuda, PT1349-018 Lisbon, Portugal, [5]Departamento de Física, Faculdade de Ciências, Universidade de Lisboa, Edifício C8, Campo Grande, PT1749-016 Lisbon, [6]School of Chemistry and Physics, Queensland University of Technology, Brisbane, QLD, Australia, [7]Centre for Astrophysics, University of Southern Queensland, West Street, Toowoomba, QLD, Australia, [8]Sydney Institute for Astronomy (SIfA), School of Physics, A28, The University of Sydney, NSW, Australia, [9]Department of Physics and Astronomy, University of Hawai'i at Mānoa, Honolulu, HI, USA, [10]Institute for Astronomy, University of Hawai'i, 2680 Woodlawn Dr., Honolulu, HI, USA, [11]INAF – Istituto di Radioastronomia, Via P. Gobetti 101, Bologna, Italy, [12]School of Natural Sciences, University of Tasmania, Private Bag 37, Hobart, Tasmania, Australia, [13]Lennard-Jones Laboratories, Keele University, Newcastle, UK, [14]E. A. Milne Centre for Astrophysics, University of Hull, Kingston-upon-Hull, UK, [15]Centre of Excellence for Data Science, AI, and Modelling (DAIM), University of Hull, Kingston-upon-Hull, UK, [16]Australian Astronomical Optics, Macquarie University, 7-9 Wally's Walk, Sydney, NSW, Australia, [17]Centre for Astrophysics, University of Southern Queensland, 37 Sinnathamby Boulevard, Springfield Central, QLD, Australia, [18]Centre for Theoretical Physics, Polish Academy of Sciences, Al. Lotników 32/46, 02-668 Warsaw, Poland, [19]School of Physics, Monash University, Clayton, VIC, Australia, [20]Department of Physics and Astronomy, University of Calgary, Calgary, AB Canada, [21]National Radio Astronomy Observatory, Socorro, NM, USA, [22]Lahore University of Management Sciences (LUMS), Lahore, Pakistan, [23]NCBC at NED University of Engineering and Technology, Karachi, Pakistan and [24]International Centre for Radio Astronomy Research, University of Western Australia, 7 Fairway, Crawley, WA, Australia

Abstract

While it is well known that galaxies are composites of many emission processes, quantifying the various contributions remains challenging. In this work, we use unsupervised machine learning based clustering algorithms to evaluate the agreement between the clustering tools and astrophysical classifications, and hence quantify the fractional contributions of star formation processes and nuclear black hole activity to the total galaxy energy budget of radio sources. We perform clustering on the multiwavelength (optical, infrared (IR), and radio) active galactic nuclei (AGN) diagnostic spaces, using the data from the G09 and G23 fields from the Galaxy and Mass Assembly (GAMA) survey, Evolutionary Map of the Universe (EMU) survey, and the Wide-field Infrared Survey Explorer. We find that the statistical clustering recovers $\approx$ 90% of the star forming galaxies (SFGs) and $\approx$ 80% of the AGN. We define a new IR-radio AGN diagnostic scheme that identifies radio AGN from IR SFGs and AGN, corresponding to the KMeans cluster with approximately 90% reliability. We demonstrate the superior power of radio AGN selection in higher dimensions using a three-dimensional space composed of directly observable parameters ($W_1 - W_2$ colour, W_2 magnitude, and the 1.4 GHz radio flux density). This novel three dimensional diagnostic shows immense potential in radio AGN selection that is close to 90% reliable and 90% complete. We also publish a catalogue of radio sources in the EMU survey with associated probabilities for them to be active in the optical regime, through which we emphasise the philosophy of considering a galaxy to be composed of various fractions rather than a binary classification of SFGs and AGN.

Keywords: galaxies: active; radio continuum: galaxies; methods: statistical

(Received 30 September 2025; revised 28 January 2026; accepted 3 March 2026)

1. Introduction

Star formation and black hole accretion are two of the most prominent energy sources that contribute to the emission from a galaxy. The interplay between these two processes is driven by an observed co-evolution of black holes and their host galaxies (Kormendy & Ho 2013), such that these processes either boost or quench each other. Signatures of these processes are ubiquitous in the spectra and colours of galaxies. Disentangling them is crucial in understanding the evolutionary properties of galaxies and in extracting important astrophysical parameters (e.g. the star formation rate (SFR) and stellar mass). Traditionally, this has been achieved using optical spectral line ratios (the BPT diagram and its variants; Baldwin, Phillips, & Terlevich 1981; Osterbrock & De Robertis 1985; Veilleux & Osterbrock 1987), infrared (IR)

Corresponding author: Jahang Prathap, Email: jahangprathap12@gmail.com.

Cite this article: Prathap J, Hopkins A, Carvajal R, Cowley M, Croom S, Farrah D, Prandoni I, Shabala S, van Loon J T, Pappalardo C, Pimbblet K, Ahmed UT, Bilicki M, Brown M, Leahy D, Mailvaganam A, Marvil J, Mukherjee T, ur Rahman S F, Vernstrom T, Willingham J and Zafar T. (2026) EMU/GAMA: A statistical perspective on active galactic nuclei diagnostics. *Publications of the Astronomical Society of Australia* **43**, e047, 1–17. https://doi.org/10.1017/pasa.2026.10175

© The Author(s), 2026. Published by Cambridge University Press on behalf of Astronomical Society of Australia. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

colours (e.g. Stern et al. 2012; Messias et al. 2012; Assef et al. 2018), radio luminosity threshold (e.g. White et al. 2000; Delvecchio et al. 2017; Magliocchetti et al. 2017; Prathap et al. 2025), X-ray luminosity threshold (e.g. Fitriana & Murayama 2022; Birchall et al. 2023), spectral energy distributions (SEDs, e.g. Cowley et al. 2018; Prathap et al. 2024), and various other multiwavelength diagnostic techniques.

The energy output of galaxies is not always dominated by star formation or black hole accretion alone (quiescent and passive galaxies are some examples); although in many cases, both contribute significantly. Traditional galaxy classification techniques approach this problem in a binary perspective, thereby classifying galaxies into one of two classes: star forming galaxies (SFG) and active galactic nuclei (AGN). A problem with this approach is that the exact astrophysics of the emission mechanisms gets hidden underneath the binary label. This can be solved by quantifying the fractional contributions from the various processes in a galaxy and is readily achieved by SED analysis, integral field studies (e.g. Xu et al. 2024), and by radio observations accurately localising the radio cores (e.g. Morabito et al. 2025). There have been attempts to quantify these contributions in terms of probabilities by identifying regions in the traditional diagnostic diagrams where there is potential mixing between the populations (e.g. Juneau et al. 2011, 2013; Yan et al. 2011; Thomas et al. 2018; Birchall et al. 2023; Morabito et al. 2025).

This work investigates whether a binary classification of galaxies as either star forming or AGN-dominated is appropriate from a statistical viewpoint using clustering analysis, with a focus on radio detections. Clustering is an unsupervised machine learning (ML) technique, a task of grouping similar objects by some measure, for instance, Euclidean distance between the data points. Broadly, existing clustering techniques can be divided into two categories: hard clustering, where each data point completely belongs to a given cluster, and soft clustering, where likelihoods are evaluated for the data points and clusters to which they belong (Yang et al. 2022). The clustering tools are also divided into different types, based on the employed principles, such as partition-based (e.g. KMeans; MacQueen 1967, fuzzy-c-means, FCM; Bezdek, Ehrlich, & Full 1984), density-based (e.g. density-based spatial clustering of applications with noise, DBSCAN; Ester et al. 1996), hierarchical (e.g. Balanced Iterative Reducing and Clustering Using Hierarchies, BIRCH; Zhang et al. 1996), statistical model-based (Gaussian mixture models, GMM; Everitt et al. 2011), and grid-based.

ML tools have been applied in various multidimensional studies, including galaxy classification in high-dimensional parameter spaces (e.g. Zhang & Zhao 2003, 2004), for star-galaxy-quasar separation (e.g. Logan & Fotopoulou 2020; Baqui et al. 2021), and AGN/SFG classification (e.g. Pennock et al. 2022, 2025; Karsten et al. 2023). In this work, we evaluate the performance of four unsupervised clustering algorithms: KMeans, GMM, FCM, and BIRCH, on multiwavelength AGN diagnostic diagrams. We choose these algorithms from a pool of clustering tools since they produced results comparable with the AGN classification. Clustering is performed in a multidimensional parameter space defined by features from the multiwavelength AGN diagnostics, and the resulting clusters are examined through their projections onto widely used two-dimensional diagnostic tools. These include optical diagnostics such as the BPT diagram (Baldwin et al. 1981), the mass–excitation (MEx) diagram (Juneau et al. 2011, 2013), the blue diagram (Lamareille 2010), and the colour-excitation (CEx) diagram (Yan et al. 2011); IR diagnostics such as the WISE[a] colour-based diagrams (Mateos et al. 2012; Assef et al. 2018) and the KI diagram (Messias et al. 2012, 2014); and the MId-infrared and RADio diagram (MIRAD; Kozieł-Wierzbowska et al. 2021).

The structure of the paper is as follows: Section 2 describes the multiwavelength dataset used in this work. Section 3 introduces the various AGN diagnostic tools and Section 4 discusses the procedure followed in clustering these diagnostic spaces. Section 5 presents the results, where we also introduce a new IR-radio AGN diagnostic tool. We discuss the findings in Section 6 and conclude in Section 7. The optical magnitudes used in this work are in the AB system, and the WISE magnitudes are in the Vega system. The WISE magnitudes have been converted to AB magnitudes only in the instances where they are combined with the K-band magnitudes (the Messias et al. 2012 diagnostic, Section 3.2). Throughout the paper we assume a flat ΛCDM cosmology with the parameters $H_o = 70$ km s-1 Mpc-1, $\Omega_M = 0.3$, and $\Omega_\Lambda = 0.7$.

2. Data

The data required for AGN diagnostics and associated clustering come from various multiwavelength surveys: Galaxy and Mass Assembly (GAMA; Driver et al. 2011, 2022) survey for optical photometry and spectra, Evolutionary Map of the Universe (EMU[b]; Norris et al. 2011, 2021; Hopkins et al. 2025) survey for the radio data, and WISE survey (Wright et al. 2010) for the IR magnitudes.

2.1. Optical data

The optical AGN diagnostics of interest in this work use spectral line intensities and photometric magnitudes, which are obtained from the GAMA survey. GAMA is a multiband imaging and spectroscopic survey undertaken at the Anglo-Australian Telescope (AAT) over five separate fields (G02, G09, G12, G15, and G23) with a limiting magnitude of $m_r < 19.8$ ($m_i < 19.2$ for the G23 field), totalling 286 deg^2. The GAMA fields other than G02 benefit from the overlap with multiple multiwavelength surveys (Bellstedt et al. 2020), such as the Galaxy Evolution Explorer (GALEX, Martin et al. 2005), European Southern Observatory (ESO) VLT Survey Telescope Kilo Degree survey (KiDS, de Jong et al. 2015), ESO Visible and Infrared Survey Telescope for Astronomy (VISTA) Kilo-degree Infrared Galaxy Public survey (VIKING, Edge et al. 2013), WISE, and Herschel (Pilbratt et al. 2010) Astrophysical Terahertz Large Area Survey (ATLAS) (Eales et al. 2010). Along with the multi-band photometry, these fields also benefit from the spectroscopic data from the AAT AAOmega spectrograph. We are interested in the G09 ($129^o < \mathrm{RA} < 141^o$, $-2^o < \delta < 3^o$) and G23 ($339^o < \mathrm{RA} < 351^o$, $-35^o < \delta < -30^o$) fields since these fields are covered by the EMU survey (see Section 2.2 below). Both of these fields also benefit from a high spectroscopic completeness of 98.5% in the G09 field and 94.2% in the G23 field (Liske et al. 2015).

GAMA data are arranged as tables in data management units (DMUs). The `gkvScienceCatv02` (Bellstedt et al. 2020) table in the `gkvInputCatv02` DMU contains the optical magnitudes, which also provides the necessary K-band magnitudes (see Sections 2.4 and 3.2). Stellar mass estimates are adopted from the `StellarMassesGKVv24` table (Taylor et al. 2011) in

[a]Wide-field Infrared Survey Explorer.
[b]https://emu-survey.org/.

https://doi.org/10.1017/pasa.2026.10175 Published online by Cambridge University Press

the StellarMassesv24 DMU. The spectral lines ([N II], [O II], [O III], Hα, and Hβ) are taken from the GaussFitSimplev05 table (Gordon et al. 2017) in the SpecLineSFRv05 DMU. We restrict to sources flagged IS_BEST, indicating reliable spectroscopic measurements. This criterion makes sure that we select the best spectrum for a given source from GAMA itself since the catalogue contains spectra from other surveys as well for comparison (see Gordon et al. 2017, for more details). The G09 and G23 fields contain 58 498 and 44 476 optical sources, respectively. This implies a total of 102 974 sources with optical spectra, photometry and stellar masses. Additional signal-to-noise (S/N) cut-off applied on the spectral lines for the optical sample is described in Section 2.4.

2.2. Radio flux densities

EMU (Norris et al. 2011, 2021; Hopkins et al. 2025) is an ongoing wide-field radio continuum survey project to deliver the touchstone radio atlas of the southern hemisphere using the Australian Square Kilometre Array Pathfinder (ASKAP; Hotan et al. 2021) radio telescope. The survey aims at covering 2π sr of the sky and cataloguing 20 million extragalactic sources, including SFGs up to redshift $z \sim 1$, powerful starbursts and AGN reaching even greater redshifts (Hopkins et al. 2025).

Our focus is on the two GAMA fields, G09 and G23, for the integrated radio flux densities. The G23 field has been observed as a part of the EMU early science project (Gürkan et al. 2022) covering an area of 82.7 deg^2 with a sensitivity of 0.038 mJy beam^{-1} at a frequency of 887.5 MHz. Out of the 55 247 sources detected, 39 812 have S/N ratios $\geq$ 5. The G09 field has been observed as a part of the EMU main project at 943 MHz and the data is internally released. The G09 field as processed by the EMUCAT pipeline (Hopkins et al. 2025, Marvil et al., in prep.) covers an area of around 96 deg^2 with 58 898 sources, where 37 981 of them have S/N$\geq$5. Since the radio diagnostic described in Section 3.3 uses the radio flux density at 1.4 GHz ($S_{1.4\,\mathrm{GHz}}$), the radio flux densities of both G09 and G23 sources are scaled to 1.4 GHz assuming synchrotron emission with a spectral index $\alpha = -0.7$, where $S_\nu = \nu^\alpha$. We do not analyse extended and compact radio sources separately. An analysis with this focus is being done by Barnes et al. (in preparation), where extended radio sources are identified and classified using ML tools.

2.3. Infrared magnitudes

WISE is an all sky IR survey that scanned the entire sky in four bands, namely W_1, W_2, W_3, and W_4 corresponding to 3.4, 4.6, 12, and 22 microns with 5σ sensitivities of 0.08, 0.11, 0.8, and 6 mJy, respectively. The W_1, W_2, and W_3 bands have better sensitivities and S/N values compared to W_4 (Wright et al. 2010), due to which we focus on the first three bands. Subsequent data releases of the WISE survey include the AllWISE (Cutri et al. 2012, 2021) catalogue consisting of 747 634 026 sources with the four magnitudes. The latest data release, the CatWISE2020 catalogue (Marocco et al. 2021) contains 1 890 715 640 sources, but with only the W_1 and W_2 magnitudes. Since the diagnostic tools that we are interested in this work require the three WISE bands, we choose the AllWISE catalogue for IR magnitudes (profile-fit photometry). The three WISE bands used for analysis (see Section 5.2) have a cut-off of S/N$\geq$5 applied (see Section 2.4 for details). The CatWISE2020 catalogue is used only for positional crossmatching as an intermediate step in identifying radio and optical counterparts (see Section 2.4), not for IR magnitudes. Accordingly, some radio sources with optical counterparts in GAMA may be lacking AllWISE counterparts and their IR magnitudes.

2.4. Final samples

We use two separate samples for clustering in optical, IR, and radio regimes. Combining all three datasets would exclude many IR and radio AGN (Leahy et al. 2019) since an AGN may not show AGN signatures in all of these bands simultaneously (Heckman & Best 2014). A solution would be using a censored dataset, where for some galaxies, certain physical attributes are missing. For instance, optical-IR-radio catalogues will have galaxies detected in IR and radio, where optical counterparts do not exist. We could model the parameters of interest for such a sample in the clustering space by masking the missing parameters. This does not mean data imputation, a statistically driven approach where the values of the missing parameters are manually inserted, because such a procedure can lead to a bias towards the properties of the sample at hand. Rather, a data censoring approach would involve constructing a marginal probability distribution without the missing parameter so as to enable clustering. However, at this stage, this procedure was deemed beyond the scope of this work (see Pennock et al. 2025 and Carvajal et al. in preparation for similar approaches), so we decided to analyse the optical and IR-radio regimes separately with an optical-radio sample and an IR-radio sample, respectively, keeping radio data as the link between the two spectral regimes.

Radio sources are often cross-matched with the optical catalogue using their IR counterparts (Norris et al. 2021, also see Anih et al. in preparation for a similar strategy selecting near-IR counterparts to X-ray sources). This is because of the strong correlation between the IR and radio emission in galaxies (Condon 1992). Also, WISE provides precise positions with respect to EMU because of its better spatial resolution. We follow this approach and as the first step in generating the optical-radio sample, we crossmatch the radio data with the CatWISE2020 catalogue (to get the coordinates of the IR counterparts of the radio sources) with a 5" radius (Norris et al. 2021), resulting in 37 981 sources in G09 and 34 841 sources in G23. CatWISE2020 catalogue incorporates a much longer time baseline relative to AllWISE, resulting in, although similar positional accuracies, but likely includes fainter sources (Marocco et al. 2021). Because of this, we use the CatWISE2020 positions instead of AllWISE for cross-matching. This radio catalogue containing the CatWISE2020 positions is then cross-matched with GAMA with a radius of 5" (Norris et al. 2021), resulting in 5 119 G09 sources and 5 890 G23 sources. WISE sources have an angular resolution of 6" in the first three bands and 12" in W_4 (Wright et al. 2010). EMU has an angular resolution of 15" (Hopkins et al. 2025) which, when compared with the WISE resolution, deems 5" a fair cross-match radius. Some radio sources are extended objects, which can result in spurious matches when a fixed crossmatch radius is used, especially when finding optical counterparts. The approach we adopted here, using IR positions for finding optical matches and the choice of an appropriate cross-match radius (Ahmed et al. 2024), helps in minimising spurious matches.

We applied quality cuts such that the various spectral line ratios used in the optical diagnostic tools ([NII]/Hα, [OII]/Hβ, and [OIII]/Hβ) have S/N$\geq$1. We expect this S/N cut to remove the radio AGN with passive spectra or undetected lines. The effects of this selection bias are discussed in Section 6.1. Certain spectral line values during the fitting procedure are replaced by the dummy

https://doi.org/10.1017/pasa.2026.10175 Published online by Cambridge University Press

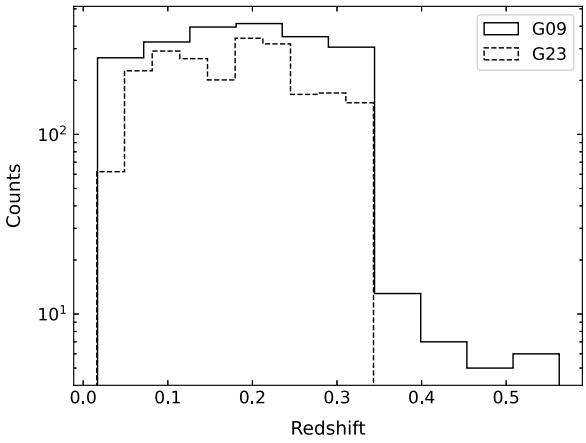


Figure 1. The redshift distributions of the radio sources in G09 (solid line) and G23 (dashed line) fields. These radio sources belonging to the optical-radio sample are selected as described in the text (see Section 2.4). The requirement of Hα line for BPT classification results in the sudden drop at $z \approx 0.34$. A few higher redshift G09 objects are most likely resulting from the higher completeness and slightly fainter magnitude limit of the field.

value '−99999.0' when there are not enough pixels to perform a fit due to bad pixels or limits of the spectral wavelength range (Gordon et al. 2017). After removing such entries from the line ratios, the final sample contains 2 091 G09 optical-radio sources and 2 193 G23 optical-radio sources. Figure 1 shows the redshift distributions of these samples, where solid and dashed lines correspond to G09 and G23 sources, respectively. The sudden drop in the redshift distribution of both G09 and G23 sources can be attributed to the requirement of Hα emission lines for the optical AGN classification (see Section 3.1). Few objects in the G09 sample extend to higher redshifts, most likely resulting from its higher completeness and slightly fainter magnitude limit. We do not apply a higher S/N, for instance, a S/N=5 threshold, on the spectral lines because this will drastically reduce the sample size. Additionally, the results described in Section 5 are not strongly dependent on this choice.

For the IR-radio sample, we crossmatch the AllWISE and radio catalogues with a radius of 5" (Norris et al. 2021). We use the AllWISE catalogue here (and not the CatWISE2020 positions as we did for optical-radio matching) because the IR diagnostic tools employed in this work require photometry in the first three WISE bands (see Section 5.2). We use the same cross-match radii while using CatWISE2020 or AllWISE catalogues because they have similar positional accuracies (Wright et al. 2010; Marocco et al. 2021). This crossmatch resulted in 32 844 G09 sources and 28 947 G23 sources. We applied a quality cut on the three (W_1, W_2, and W_3) WISE bands so that the magnitudes satisfy S/N≥5, resulting in 14 123 G09 IR-radio sources and 14 339 G23 IR-radio sources. It should be noted that the KI diagram described below in Section 3.2 requires having a K-band magnitude. This restricts those results to only a much smaller number of sources (3 755 in G09 and 4 101 in G23). Because of this, we do not include the K-band magnitudes as an additional parameter in the clustering analysis. We also do not apply any S/N criterion on K-band magnitudes.

3. AGN classification

We employ a suite of multiwavelength spectral and photometric AGN diagnostics in this work, comprising four optical diagnostics, three IR colour-magnitude diagnostics, and one IR-radio flux-based diagnostic. We do not employ X-ray data in this analysis, as a crossmatch of the relevant samples with the X-ray data (first eROSITA all-sky survey (eRASS1); Merloni et al. 2024) resulted in a sample of only a few hundred galaxies, which would preclude the clustering analysis. This section details the relevant diagnostic tools and our approach in employing them for our purposes. These diagnostics are visualised in Figures 2 and 3. The sources in the respective samples are classified using these diagnostics in a binary fashion, and these labels are used only for comparison with the clustering outputs described in Section 4.

3.1. Optical diagnostics

3.1.1. BPT diagram

The Baldwin, Phillips, and Terlevich (BPT) diagram saw its inception in the work done by Baldwin et al. (1981). The BPT diagram uses the spectral line ratios [NII]/Hα and [OIII]/Hβ for ionising source classification into AGN and SFGs, and is used extensively (e.g. Prathap et al. 2024, 2025; Hermosa Muñoz et al. 2025; Stevenson et al. 2025). Further improvements were proposed by Kewley et al. (2001) for a theoretical maximum starburst line and by Kauffmann et al. (2003) for a semi-empirical classification (also see, Osterbrock & De Robertis 1985; Veilleux & Osterbrock 1987). The BPT diagram, as used in this work (see Figure 2a), classifies galaxies into three categories: (*i*) SFGs, if they fall below the Kauffmann et al. (2003) line, (*ii*) AGN, if they lie above the Kewley et al. (2001) line, and (*iii*) composites, if they fall in between the two demarcation lines.

3.1.2. MEx diagram

The mass-excitation (MEx) diagram was proposed by Juneau et al. (2011) as an alternative to traditional nebular line diagnostics as the [N II] and Hα lines are redshifted out of the optical spectra for sources at high redshifts (say, for $z > \approx 0.5$). The MEx diagram uses [OIII]/Hβ and galaxy stellar mass as a proxy for [NII]/Hα. The choice of stellar mass instead of [NII]/Hα can be justified using the mass-metallicity relationship (Tremonti et al. 2004). The MEx diagram classifies galaxies into SFGs, AGN, and composites (see Figure 2b). In this work, we use the demarcation lines defined in Juneau et al. (2013).

3.1.3. Blue diagram

Galaxy classification has also been explored using only the blue spectral lines ([O III], [O II], and Hβ) to enable the analysis of high redshift sources. Lamareille et al. (2004) introduced the blue diagram which uses the line ratios [OII]/Hβ and [OIII]/Hβ and was improved by Lamareille (2010). The diagram contains five regions corresponding to SFGs, AGN, low-ionisation nuclear emission line regions (LINERs), composites of SFGs and AGN, and the composites of SFGs and LINERs. The true nature of LINERs is actively debated where spatially resolved studies (e.g. Belfiore et al. 2016) have identified LINER-like spectra in star-forming and quenched galaxies. Márquez et al. (2017) demonstrated that a large fraction of the LINERs in their sample of X-ray sources can be considered genuine AGN. In this work, we classify LINERs as AGN, thus reducing the composite class to a single population, resulting in a three-system classification like the BPT and MEx classifications (see Figure 2c).

https://doi.org/10.1017/pasa.2026.10175 Published online by Cambridge University Press

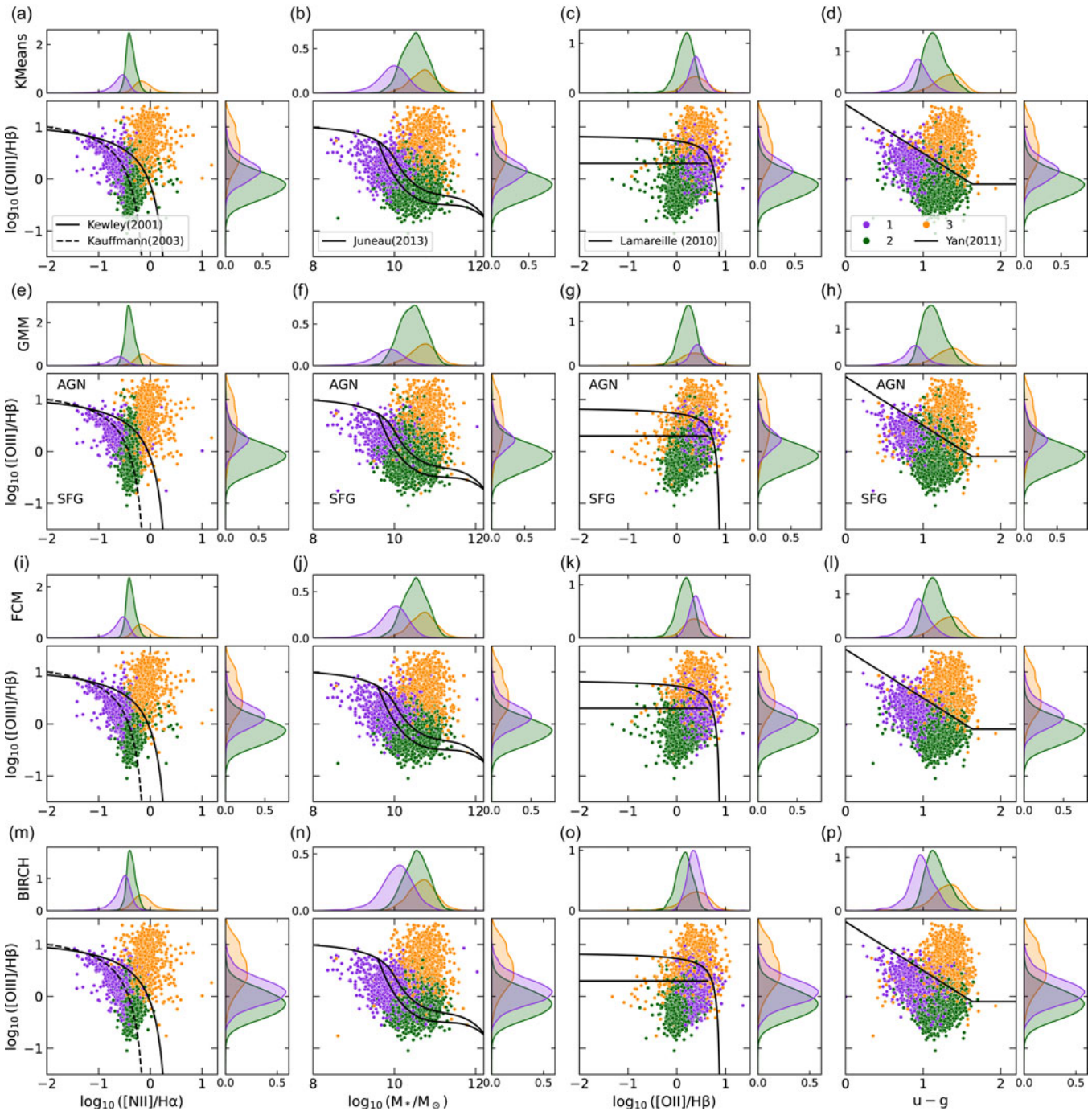


Figure 2. Performance of various clustering tools in different optical diagnostic spaces. The figure shows the different clustering tools row-wise, and the different empirical diagnostic tools are presented along the columns. Panels a–d: the clusters identified by KMeans are plotted in the optical diagnostic spaces BPT diagram, MEx diagram, blue diagram, and CEx diagram, respectively. Panels e–h: the clusters identified by GMM, panels i–l: the clusters identified by FCM, and panels m-p: the clusters identified by BIRCH are plotted in the same order as the empirical diagnostics. The SFG and AGN regions are labelled in the second row. Composite galaxies in these diagnostic plots occupy the region between the demarcation lines in the first three columns, the CEx diagram does not define a composite region. The purple and green clusters correspond to the star forming species, following a metallicity sequence. The orange clusters seemingly occupy the region identified as AGN in each of these diagnostic spaces. Each of these plots features two marginal histograms showing the normalised densities corresponding to the three clusters identified by each of the clustering tools.

3.1.4. CEx diagram

The colour-excitation (CEx) diagram (Yan et al. 2011) follows a similar approach to the MEx and the blue diagrams, to be able to classify high redshift galaxies. The CEx diagram is inspired by the correlation between rest frame H-band magnitude and the metallicity for SFGs (Weiner et al. 2007; Yan et al. 2011). This relation collapses for AGN and thus the $U-B$ colour, along with [OIII]/Hβ, was proposed as a more effective probe for galaxy classification. The CEx diagram classifies galaxies into two classes, SFGs and AGN; a composite region is not defined in this case (see Figure 2d). We used the K-correction relations in

https://doi.org/10.1017/pasa.2026.10175 Published online by Cambridge University Press

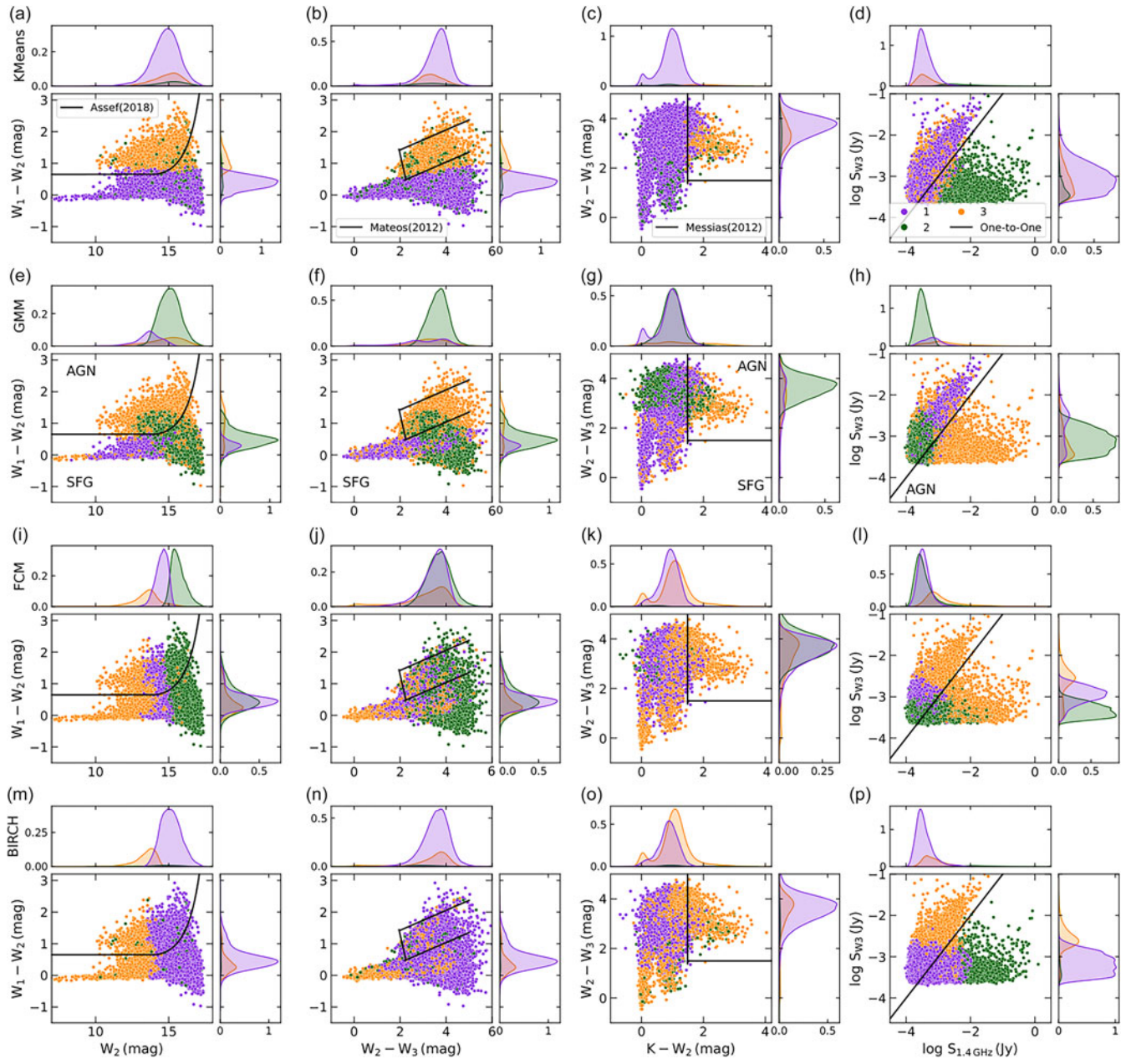


Figure 3. Performance of various clustering algorithms in different IR and radio diagnostic spaces. The figure shows the different clustering tools row-wise, and the different empirical diagnostic tools are presented along the columns. Panels a–d: the clusters identified by KMeans are plotted in the IR diagnostic spaces defined by Assef et al. (2018), Mateos et al. (2012), Messias et al. (2012), and the IR-radio diagnostic space defined by Kozieł-Wierzbowska et al. (2021), respectively. Panels e–h: the clusters identified by GMM, panels i–l: the clusters identified by FCM, and panels m-p: the clusters identified by BIRCH are plotted in the same order as the empirical IR-radio diagnostics. The clustering seems to be working well only in the case of KMeans (a–d), where we are able to compare the clusters and the empirical classifications. In the panels a–d, the purple cluster represents IR SFGs, the orange cluster represents IR AGN, and the green cluster represents radio AGN (see text for details). Each of these plots features two marginal histograms showing the normalised densities corresponding to the three clusters identified by each of the clustering tools.

Blanton & Roweis (2007) to match the $U - B$ criterion of Yan et al. (2011) to the $u - g$ colour generated using GAMA data.

3.2. Infrared diagnostics

AGN and star formers show significant differences in their IR SEDs (e.g. Lyu & Rieke 2022). This feature has been exploited in the form of near- and mid-IR colour criteria in developing IR AGN diagnostic diagrams. Multiple surveys have covered the sky in IR. These include the Two-Micron All Sky Survey (2MASS; Skrutskie et al. 2006), Spitzer Space Telescope (Werner et al. 2004) using the infrared array camera (IRAC; Fazio et al. 2004), and WISE. The AGN selection criteria developed using the data from these surveys include, the Lacy wedge (Lacy et al. 2004, 2007; Sajina, Lacy, & Scott 2005), the Stern wedge (Stern et al. 2005), the IRAC power-law (Alonso-Herrero et al. 2006; Donley et al. 2007, 2012), the KI & KIM methods (Messias et al. 2012, 2014), and the WISE selections (Stern et al. 2012; Assef et al. 2013, 2018). In this

https://doi.org/10.1017/pasa.2026.10175 Published online by Cambridge University Press

work, we focus on the three diagnostic tools that can be built using WISE magnitudes.

Stern et al. (2012) put forward a magnitude-independent AGN selection criterion that selects AGN reliably up to W_2 magnitude brighter than 15.05 mag. As an improvement, Assef et al. (2018) proposed a magnitude-dependent selection, with $W_1 - W_2$ colour as a function of W_2 magnitude, that selects AGN reliably up to W_2 magnitudes brighter than 17.11 mag. The demarcation line in this diagram varies depending on the reliability and completeness of the resulting AGN sample. In this work, we use the version optimised for reliability (equation 4 of Assef et al. 2018), that classifies the galaxies into SFGs and AGN (see Figure 3a).

Mateos et al. (2012) defined a wedge that primarily selects X-ray AGN that are also active in the IR. They proposed two versions of the diagram, where the first version uses the first three WISE magnitudes, and the uses all four WISE magnitudes. We choose the former in this work, because the W_4 band is relatively shallow and may only select bright objects. The three-band wedge uses the colours $W_1 - W_2$ and $W_2 - W_3$ for delineating the SFGs and AGN (see Figure 3b).

The KI diagram proposed by Messias et al. (2012) uses the K and IRAC bands (K-IRAC colour). IRAC bands facilitate AGN identification since they capture the polycyclic aromatic hydrocarbon (PAH) features in the mid-IR bands. In the EMU fields considered in this work, however, we do not have IRAC data available. Instead, we replace the IRAC bands in the KI diagram with the WISE bands. The 4.5 micron IRAC band is close to the W_2 band (4.6 microns) in wavelength. Instead of the 8 micron IRAC band, we use the WISE W_3 band (12 microns). Both WISE bands are converted to AB magnitudes in this case since the Messias et al. (2012) criteria are defined in the AB system. We acknowledge the differences in the widths of the IRAC and WISE bands, and that they might probe additional features than what the corresponding IRAC bands do. As evident in Section 5, this difference in wavelength does not drastically affect the classification. The Messias et al. (2012) diagnostic also classifies the galaxies into SFGs and AGN. There are no composite classes identified in the IR diagnostics used in this work (see Figure 3c).

3.3. Radio diagnostics: The MIRAD diagram

Galaxies with radio emission driven by black hole accretion are termed radio AGN. Various classes of such objects exist, and multiple definitions based on specific use cases are employed while selecting radio AGN (see Heckman & Best 2014, for a review on AGN and different types of radio AGN). The most prominent radio AGN selection uses the FIR–radio correlation (Condon 1992), a relation tightly followed by SFGs. Sources that deviate significantly from this correlation are interpreted as radio loud AGN. An upper limit on the radio luminosity (White et al. 2000; Prathap et al. 2025), and radio-to-optical flux ratio (Kellermann et al. 1989) are also widely used techniques for radio loud AGN selection, hereafter, termed radio AGN. As discussed in Prathap et al. (2025), the scarcity of ample radio sources with spectroscopic counterparts restricts the usage of luminosity as a classification criterion in many cases. An alternative to this approach is the flux-flux diagrams with an appropriate selection criterion (similar to flux-ratio methods).

Given these considerations, in this work we use the MId-infrared and RADio (MIRAD; Kozieł-Wierzbowska et al. 2021) flux-based diagram for radio AGN selection (see Figure 3d). It uses the WISE W_3 flux, $S_{1.4\,\mathrm{GHz}}$, and a simple one-to-one line as a selection criterion. In this diagram, pure star formers occupy the upper sequence and AGN, since they generally show excess radio emission for a given mid-IR flux, occupy the lower sequence. It is interesting to note that recent studies using LOw-Frequency ARray (LOFAR) (e.g. Morabito et al. 2025) have shown that many compact sources without excess radio emission also host AGN-related radio emission. In these perspectives, this selection is potentially a lower limit on the radio AGN numbers.

4. Clustering

A wide range of clustering algorithms is available, and assessing their performance in each diagnostic space is essential. In this work, we use the KMeans, GMM, FCM, and BIRCH algorithms to statistically classify the radio galaxies. These four tools are chosen since they represent different clustering principles, so that the performance of both the clustering principles (see the subsections below for a description of the algorithms and the principles employed by them) and the tools can be assessed. The input to the clustering tools is a set of physical observables used in the diagnostic diagrams discussed earlier, resulting in an n dimensional space, where the clustering is realised. Analyses in higher dimensional space, as we encounter here, can be supported by dimensionality reduction techniques like principal component analysis (PCA; Hotelling 1936) where we can choose the components carrying the most variance among the whole input parameter set. Clustering can then be done on these principal components with the reduced dimensional complexity. As we show later in Section 5, PCA is an optional step, and whether it supports the clustering is dependent on the specific spectral regime. All parameters used for clustering are standardised using the `RobustScaler` method in the `Python` package `scikit-learn` (Pedregosa et al. 2011) to a mean of zero and standard deviation of one. This is important, especially in astronomical data, since most parameters scale logarithmically. In the rest of this section, we give an overview of the different algorithms used.

4.1. KMeans

KMeans belongs to the broad category of centroid (partition)-based clustering (Yang et al. 2022). It is a simple and robust algorithm, iteratively classifying the data into a predefined number of clusters by minimising the total sum of the distances between the data points and the centres of the respective clusters (Fotopoulou 2024). KMeans is used extensively in astronomy for multiple applications, such as spectral classification and improving S/N values by stacking the spectra (see, Ordovás-Pascual & Sánchez Almeida 2014, and references therein). The two main decisions one has to make while initialising KMeans are (*i*) the threshold on the distances between the data points and (*ii*) the number of cluster centroids (k), which in turn decides the number of clusters. Each point is assigned to the cluster centred at the closest centroid based on the defined distance metric. This process is repeated, taking the mean positions of the points associated with a given cluster until convergence. Thus, the choices initially made can have a great impact on the outcome and the convergence time (Baron 2019).

Choosing the best k value is non-trivial. It is usually done using the elbow method, which minimises the inertia, a measure of the

https://doi.org/10.1017/pasa.2026.10175 Published online by Cambridge University Press

sum of the squared distances of each data point to its closest cluster centre (also see, Sánchez Almeida et al. 2010). Another measure is the silhouette coefficient, which quantifies how far away the data points are from the neighbouring cluster. We use the `KMeans` function from the `Python` package `scikit-learn` to perform the KMeans clustering in our analysis. We used both the elbow method and the silhouette coefficient method to identify the best k value, which turned out to be either three or four. As described below in Section 5, we choose three as the value of k, the number of clusters, in the case of all the other clustering tools discussed.

4.2. Gaussian mixture models

GMMs are parametric probability density functions represented as a weighted sum of Gaussian probability densities. GMM, as a clustering tool, belongs to the general statistical model-based clustering methods, which models the observed data distribution as a random variable drawn from an underlying multivariate distribution (de Souza et al. 2017; Fotopoulou 2024, and references therein).

GMM also requires the number of clusters to be predefined, but it is a soft clustering method, meaning every data point will be assigned a probability for it to belong to the respective cluster. Various fitting methods are available, but the expectation-maximisation (EM) algorithm (Dempster, Laird, & Rubin 1977) is widely used for finding the optimal clusters. In this work, we use the `GaussianMixture` function from `scikit-learn` for our analysis. One of the parameters that we adjusted was the `covariance_type`. This parameter decides the covariance matrix of the Gaussian components, and hence the different shapes the clusters can adopt. We choose the `covariance_type` of the `GaussianMixture` function to be 'diagonal' since this results in cluster distributions similar to other clustering tools. A choice of any of the available `covariance_types` results in similar cluster density distributions with only small rearrangements in the number of data points in the metallicity sequence (see Section 5.1).

4.3. Fuzzy C-means

Fuzzy sets as a class of objects with a continuum of membership grades was introduced by Zadeh (1965). Since then, clustering problems based on fuzzy sets have been extensively studied (see, Nayak, Naik, & Behera 2015, and references therein). Clustering algorithms that minimise an objective function and assign classes using the fuzzy technique are generally called fuzzy C-means (FCM; Bezdek et al. 1984; Nayak et al. 2015). FCM belongs to the set of partition-based methods, similar to KMeans in its clustering rationale, but in a fuzzy way, assigning probabilities (or the degree to which a given data point belongs to the clusters) to data points on their cluster assignments. FCM also require the number of clusters to be predefined. We use the `Python` package `fuzzy-c-means` (Dias 2019) in this work. The function has a parameter m corresponding to the degree of fuzziness, which quantifies the degree to which each datapoint belongs to a given cluster component. For $m > 2$, we found that the degree of fuzziness gets constrained between 0.3 and 0.7. This is not desirable since we want to explore the entire 0–1 range, as this range corresponds to a measure similar to a probability (although note that the degree of fuzziness parameter is not a probability). So we choose a value of $m = 1.4$. This choice has implications when we calculate the probabilities in Section 6.2.

4.4. BIRCH

BIRCH (Zhang et al. 1996) is a hierarchical clustering algorithm built particularly for large datasets. Its main feature is that each clustering decision is made locally, that is, without scanning the entire dataset. This enables high speed clustering, especially beneficial for large datasets. It does so by using the characteristics of the clustering feature (CF) tree (a compact representation of a cluster based on data statistics, see Zhang et al. 1996, for more details). The choice of the number of clusters and a parameter called `threshold` plays a crucial role while using BIRCH. The parameter `threshold` controls the resolution of each CF tree, where it acts as an upper limit for the CF tree components to split. These components are merged to form the desired number of clusters based on the predefined number of clusters. We use a value of 0.1 for `threshold`, since lower values promote splitting of the clusters, and have resulted in cluster shapes comparable with astrophysical classifications (see Section 5.1). We use the `BIRCH` function from `scikit-learn` in `Python` to perform clustering.

5. Results

The eight diagnostic diagrams described in Section 3 are investigated for clustering with the tools introduced in Section 4 and the results are presented in this section. In the discussions below, clusters refer to the groupings identified by the clustering tools and classes (or classifications) refer to the astrophysical classifications assigned by the AGN diagnostic tools. We tried various input parameter combinations for the clustering tools (different combinations of fluxes, magnitudes, colours, and flux ratios), and the optimal choice proved to be flux ratios and colours in the optical diagnostics and magnitudes (and radio fluxes) in the IR-radio diagnostics. This is not surprising since these parameters represent the spaces where the data points exhibit segregation. We choose three as the number of clusters for better comparability with the astrophysical classifications. We find, using the functions associated with KMeans (the elbow method and the silhouette coefficient method), that the optimal number of clusters is either two or three, reinforcing our choice.

An optional step of dimensional reduction using PCA can be done before clustering. We investigated the effect of PCA on the clusters identified by different algorithms. This proved to be dependent on the respective spectral regime. The clusters identified in the optical diagnostic spaces compare well with the astrophysical classifications when the parameters do not undergo dimensional reduction. The clusters identified in the IR-radio spaces, on the other hand, compare well with the astrophysical classifications, if the principal components of the parameters are used. Essentially, optical-radio data do not benefit from PCA, whereas IR-radio data require PCA. It is not clear why this is the case. It could be because of the inherent differences in the optical and IR diagnostic parameter spaces, but exploration of this aspect is beyond the scope of the current work. PCA, when done, uses the `Python` package `pca` (Taskesen 2020), where the principal components carry at least 99% of the variance. As discussed in Section 2, the optical-radio and the IR-radio data undergo clustering separately in their respective parameter spaces and these clusters are projected to the two-dimensional diagnostic spaces for analysis.

5.1. Optical-radio clusters

Figure 2 shows the clusters identified by the different algorithms in the optical-radio sample in a suite of optical diagnostics. The four

https://doi.org/10.1017/pasa.2026.10175 Published online by Cambridge University Press

panels show the clustering done by KMeans (Figure 2a–2d), GMM (Figure 2e–2h), FCM (Figure 2i–2l), and BIRCH (Figure 2m–2p), respectively. In every row there are four panels, each showing an optical diagnostic in the order BPT, MEx, Blue, and CEx diagrams, respectively. For clustering these spaces, we choose six parameters: log([OIII]/Hβ), log([OII]/Hβ), log([NII]/Hα), stellar mass (log M_*), $u-g$ colour, and log $S_{1.4\,\mathrm{GHz}}$. The resulting clusters are then projected onto the two dimensional optical diagnostic spaces. These parameters are chosen as the combinations used in the relevant AGN diagnostics with the radio flux density as an additional constraint. Each clustering tool identifies three clusters distinguished by their colours: purple, green, and orange. The marginal plots show the density of each of these clusters for the corresponding axis parameter. No PCA was applied on the parameters for optical diagnostics since a dimensional reduction seemed less effective compared to the results presented here.

Upon visual inspection, it is evident that the three clusters identified by the different algorithms (along each row), as shown in Figure 2, roughly follow the astrophysically motivated classifications (along each column). For instance, in Figure 2(a), we can see a clear distinction of the purple and green clusters into a star forming set towards smaller [NII]/Hα and below the demarcation lines specified by Kewley et al. (2001) and Kauffmann et al. (2003), and of the orange cluster into an AGN set above the Kewley et al. (2001) line. The clusters that likely represent the SFG population is divided into two (purple and green dots), which most likely correspond to a metallicity sequence where the purple cluster represents low-metallicity star formers and the green cluster represents the high-metallicity star formers. This is supported by the separation of these clusters as a function of galaxy stellar mass (the mass-metallicity relation; Tremonti et al. 2004), in the MEx diagram shown in panels 2(b), 2(f), 2(j), and 2(n). This split of the SFG population into a metallicity sequence is most likely driven by the data distribution in the six dimensional parameter space used for clustering.

Even though clusters are distinguishable when compared to the demarcation lines, they are not completely separate, as evident from the marginal density distributions. We do not expect both statistical clusters and astrophysical classifications to be identical since both techniques treat the parameter space uniquely. The overlap between the multiple clusters, shown consistently throughout the clustering tools, might be an indication of a softer boundary between the populations, a concept we explore throughout in this work.

5.2. Infrared-radio clusters

Figure 3 shows three IR AGN diagnostics, in the first three columns, and the radio AGN diagnostic in the last column. It shows the action of KMeans (Figure 3a–3d), GMM (Figure 3e–3h), FCM (Figure 3i–3l), and BIRCH (Figure 3m–3p), respectively, on these diagnostic spaces. Clustering is done on a four-dimensional IR-radio space containing the three WISE magnitudes, W_1, W_2, W_3 and log $S_{1.4\,\mathrm{GHz}}$. We do not use K-band magnitude as a clustering parameter. As described in Section 2.4, only around 30% of the IR-radio sample has K-band magnitude available. Thus, including the K-band magnitude as a parameter may bias the clustering results. The clustering rationale and the figure characteristics are similar to Section 5.1 where the three clusters are distinguished by their colours: purple, green, and orange. The clustering in this case, however, is done on the principal components carrying 99% variance of the astrophysical features, as they showed clearer separation.

It is evident from Figure 3 that the KMeans clusters are well distinguishable and populate the diagnostic space similar to the astrophysical classifications in the IR-radio diagnostics. The purple and orange clusters occupy the IR SFG and AGN regions, respectively, in the IR diagnostics (Figure 3a–3c). Notably, we see that the orange clusters occupy predominantly the same regions (above the Assef et al. 2018 line, within the Mateos et al. 2012 wedge, and the Messias et al. 2012 wedge in columns a, b, and c, respectively) as those of the AGN classes from the diagnostics. The IR diagnostics considered here do not have a composite classification, and the green cluster is dispersed through the former two clusters. It is interesting to see that this dispersed population, however, occupies the region identified as AGN in the radio diagnostic in Figure 3(d). This feature is quite important, with the potential to explore radio AGN those are not flagged as IR AGN. We explore this in more detail below in Section 5.3. Both purple and orange clusters (SFGs and AGN, respectively, in the IR diagnostics) occupy the SFG region of the radio diagnostic in Figure 3(d). This means that the orange cluster represents a radio-quiet IR active galaxy population. It is clear that GMM, FCM, and BIRCH perform poorly in identifying clusters that can be compared with the astrophysical classes. This is evident from the Figure 3(e)–3(p) since the different clusters occupy regions that are not differentiated by any demarcatio+n lines of the relevant IR-radio diagnostics.

Both the first and third columns in Figure 3, corresponding to the Assef et al. (2018) and Messias et al. (2012) diagnostics, feature an extended tail towards $W_2 \lessapprox 12$ and $W_1 - W_2 \approx 0.2$. These regions in these parameter space are predominantly occupied by star-like objects, and a subsample was visually inspected to assess the stellar proportion. There are only 303 (out of 28 462, 1%) objects satisfying these criteria. For visual inspection, we cross-matched these 303 objects with the GAMA database at a 5" radius, resulting in 202 GAMA objects. For the remaining sources, we used the archival data from the Barbara A. Mikulski Archive for Space Telescopes (MAST).[c] It was found that there are star-like objects in our sample in this region. However, many objects here are also found to be extended sources. Excluding stellar objects using a selection criterion like a lower bound on redshift is not possible in this case since we do not have redshift information available for many sources in the IR-radio sample. Using a magnitude cut-off would also result in the removal of true galaxies. As we are more interested in the green cluster (see next section) and there are only 16 of these in the star-like region (of which none appear point-like), we do not remove these objects from the analysis. The small numbers involved ensure their inclusion does not significantly affect our results or conclusions.

5.3. An IR-radio AGN diagnostic

Out of the three populations identified by KMeans in Figure 3(a), the purple and orange clusters (representing IR SFGs and AGN) lie below and above the demarcation line, respectively. The green cluster is distributed broadly across the parameter space. This scatter is not clear from the marginal density plots because the green cluster members correspond to only 6% of the total population,

[c] mast.stsci.edumast.stsci.edu.

https://doi.org/10.1017/pasa.2026.10175 Published online by Cambridge University Press

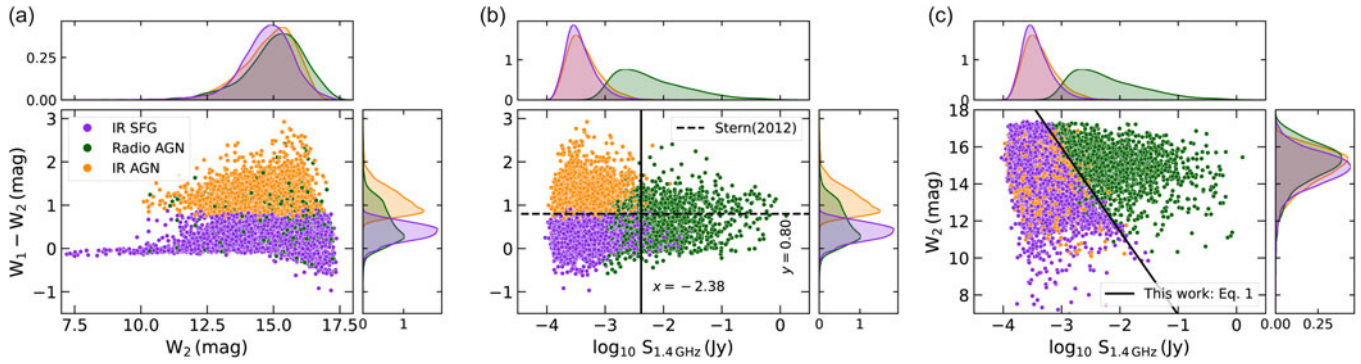


Figure 4. The distribution of KMeans clusters in various IR-radio spaces. Panel a shows the Assef et al. (2018) diagnostic space without the demarcation line, panel b shows the variation of the $W_1 - W_2$ colour as a function of $S_{1.4\,\mathrm{GHz}}$ with the dashed line representing the Stern et al. (2012) demarcation between IR SFGs and IR AGN and the solid line at $\log_{10} S_{1.4\,\mathrm{GHz}} = -2.38$ Jy separates the radio AGN from IR sources (see text for details). Panel c shows the W_2 magnitude as a function of $S_{1.4\,\mathrm{GHz}}$, where the solid line (Equation 1) separates the radio AGN from other sources. This plot is different from the Kozieł-Wierzbowska et al. (2021) diagnostic since they use the W_3 flux. The colour scheme follows Figure 3, but we are explicitly defining the purple cluster as IR SFGs, the orange cluster as IR AGN, and the green cluster as radio AGN, based on the characteristics evident from the discussions so far. The non normalised density of each of these clusters are shown as marginal densities following the same colour scheme.

and the density plots are normalised with respect to the total number of sources. It was noted in Section 5.2 that the green cluster that appears dispersed in the IR diagnostic space (Figure 3a) populates the AGN space in the radio diagnostic, as shown in Figure 3(d). To investigate if there exists any natural grouping between these clusters, we analyse a combination of the IR-radio parameter spaces using the KMeans clusters, shown in Figure 4. The colours of the clusters follow those in Figure 3(a)–3(d), where purple clusters represent IR SFGs, orange clusters represent IR AGN, and green clusters represent radio AGN, all identified using KMeans as described in Section 5.2. The marginal density plots are not normalised with respect to each other to better aid the visual distinction of the clusters.

Figure 4(a) reproduces the Assef et al. (2018) diagnostic (as shown in Figure 3a). Figure 4(b) shows the $W_1 - W_2$ colour as a function of $S_{1.4\,\mathrm{GHz}}$ with the dashed line (Stern et al. 2012) separating the IR SFGs from the IR AGN. Above the solid black line at $\log_{10} S_{1.4\,\mathrm{GHz}} = -2.38$ Jy sources are identified as radio AGN (green dots) with only 10% contamination from IR detected sources. It is well-known that for flux-densities brighter than a few mJy at this frequency the radio source population is dominated by AGN (e.g. Seymour et al. 2008).

Figure 4(c) is a variant of the diagnostic defined by Kozieł-Wierzbowska et al. (2021), also shown in Figure 3(d), and is of special interest to us. Instead of the W_3 flux, it looks at the variation of the W_2 magnitude (since we use the $W_1 - W_2$ colour and we can benefit from a sensitivity better than the W_3 magnitude) as a function of $S_{1.4\,\mathrm{GHz}}$. Notably, we see that the combination of these two parameters is able to separate the radio population from the IR-detected sources. The solid black line, defined by

$$W_2 = m\ \log_{10} S_{1.4\,\mathrm{GHz}} + c, \tag{1}$$

where the slope, $m = -4.519$ and the y-intercept, $c = 2.409$, separates the IR and radio population, such that the radio AGN lying above the line defined in Equation 1 is 90.26% reliable and 91.25% complete. It is important to note the caveats here. Being a flux-cut, it is likely that the numbers stated are specific to this sample and survey. A more sensitive radio survey can, for the same threshold, identify sources with varying reliability and completeness. Also, the completeness and reliability values reported here are based only on radio AGN. These values will increase slightly if we account for both IR and radio AGN.

5.4. A three-dimensional perspective

The previous section introduced a new IR-radio AGN selection, Equation 1, which is relatively reliable and complete. In this section, we build on this two-dimensional diagnostic with a three-dimensional approach.

Figure 5 shows a three-dimensional space composed of the Assef et al. (2018) (Figure 3a) parameters as the x- and y-axes, and $\log_{10} S_{1.4\,\mathrm{GHz}}$ as the z-axis. Figure 5(a) and 5(b) show the different orientations of the same parameter space. The data points represent the KMeans clusters identified in the Figure 3(a)–3(d), where the purple clusters represent IR SFGs, orange clusters represent IR AGN, and the green clusters represent radio AGN. The (non-normalised) densities of each of these clusters are shown on the white surfaces with projections mapped to the same colours as the data points. A clear separation of the different populations is apparent in these diagrams.

We define planes of demarcation between the different clusters in Figure 5. The grey plane separates the green clusters (corresponding to radio AGN) from the others (representing IR galaxies) and is defined as

$$a\ W_2 + b\,(W_1 - W_2) + c \log_{10} S_{1.4\,\mathrm{GHz}} + d = 0, \tag{2}$$

where the coefficients, $a = 0.4598$, $b = -0.2491$, $c = 2.3541$, and $d = -0.2201$ are chosen such that the contamination by the clusters representing the IR sources above this plane is constrained to a maximum of $\approx 10\%$. This criterion selects radio AGN such that in the region where Equation 2 > 0, the selection is 92.22% reliable and 91.49% complete. A second criterion that separates the clusters representing the IR SFGs and IR AGN is shown as the crimson plane, which is simply a two-dimensional version of the Stern et al. (2012) criterion. The high completeness and reliability of the radio AGN selection in the three-dimensional criterion, while maintaining the IR SFG-AGN separation, demonstrates the significant potential of higher-dimensional diagnostics for AGN classification.

https://doi.org/10.1017/pasa.2026.10175 Published online by Cambridge University Press

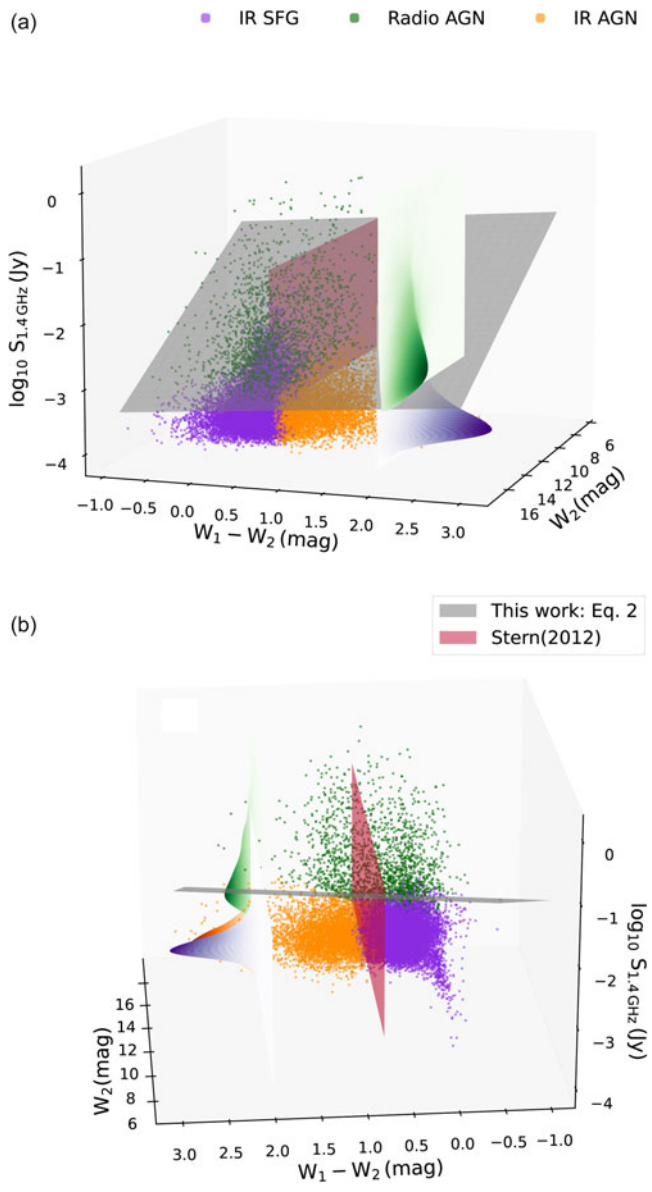


Figure 5. A three-dimensional IR-radio AGN diagnostic, combining the parameters shown in Figure 4. The colour scheme of the data points follows Figure 4, where the purple cluster represents the IR SFGs, the orange cluster represents the IR AGN, and the green cluster represents the radio AGN. The non-normalised densities of these species are shown following the same colour scheme. The grey plane, defined by Equation 2, separates the radio AGN from the IR sources with the resultant radio AGN being both complete and reliable over 90% (see text for details). The crimson red plane corresponds to the three-dimensional version of the Stern et al. (2012) criterion separating the IR SFGs and IR AGN.

The parameter space in Figure 5 is essentially a combination of the best features of Figure 4(b) and 4(c). There are multiple benefits to such a selection criterion (Equations 1 and 2), (*i*) it encodes information from two spectral regimes, namely, IR and radio, a benefit of the clustering approach; (*ii*) since these criteria do not need optical counterparts for comparison, and the IR counterparts of the radio sources are more commonly identified, the sample for AGN identification is no longer constrained by its size, thereby increasing the number of available candidates; (*iii*) the separation of IR sources into SFGs and AGN is still evident; (*iv*) a stricter radio selection resulting in radio sources where the nuclear black hole activity most likely contributes to a larger fraction of the radio emission; and (*v*) enables the selection of sources that are likely active in one spectral regime and not the other, for instance, radio AGN those would otherwise be identified as IR star formers. An implication of the last point is that the diagnostic has the potential for selecting low excitation radio galaxies (LERGs; Heckman & Best 2014; Prathap et al. 2024), a class of AGN that are active primarily in the radio.

Figure 6(a) and 6(b) test the last two points. They use the optical spectra and redshifts of the IR-radio sample (where available) to analyse IR-radio correlation in Figure 6(a), and the low-excitation nature of these galaxies, as prescribed by Best & Heckman (2012) in Figure 6(b). This sample of 9 027 sources (32% of the IR-radio sample) is constructed by cross-matching the IR-radio sample with the GAMA spectroscopic data with a 5" radius. The dotted line in Figure 6(a) is 0.7 dex (5 in linear scale) above the one-to-one (solid) line (see e.g. Delvecchio et al. 2017; Prathap et al. 2024, for a similar selection of radio-excess AGN, but with a lower threshold value of three in the linear scale). This is a conservative threshold so that the radio AGN selection minimises contamination. Around 88% of the green dots (radio AGN) lie above this line, supporting the point that AGN activity contributes to a large fraction of the radio emission. It is also worth noting that the IR-radio correlation evolves with redshift (see Cowley et al. 2016). This evolution would not affect our results since the effect is more prominent for $z > 1$ (Delvecchio et al. 2021). LERGs are expected to lie below the solid line in Figure 6(b) (see Best & Heckman 2012 for more details). It is clear that a large part of the radio AGN population lies above this line. This implies that the sources detected in optical, IR, and radio regimes are not predominantly LERGs. This is expected, as the requirement of optical spectra biases against optically weak sources, potentially removing LERG-like objects. A consensus on the LERG population seems not feasible without more sensitive spectroscopic measurements.

Mid-IR AGN selection techniques, especially comprehensive approaches such as those of Stern et al. (2012) and Assef et al. (2018), are highly effective at identifying clean samples of luminous AGN. However, they are biased against low-luminosity sources, particularly at higher redshifts (Donley et al. 2012; Mateos et al. 2012; Rovilos et al. 2014). Consequently, their results cannot be straightforwardly generalised to the full AGN population. Here, our novel approach stands out. A selection, combining mid-IR and radio observations like ours, can be used to partly rectify this bias, but deeper observations are needed for a better understanding.

6. Discussion

6.1. Clusters and classifications

6.1.1. Optical parameter space

Table 2 evaluates the performance of the various clustering tools on different optical AGN diagnostics. It is clear from the discussions so far that the purple and green clusters from the clustering tools represent the SFG class and the orange clusters represent the AGN class from the astrophysical classifications. Accordingly, we label the purple and green clusters as SFGs and the orange clusters as AGN in the discussions in the current and following sections. In Table 2, the composite class from the diagnostic tools and the metallicity sequence identified by the clustering tools (purple + green) are labelled as their respective SFG classes. That is, the SFG

https://doi.org/10.1017/pasa.2026.10175 Published online by Cambridge University Press

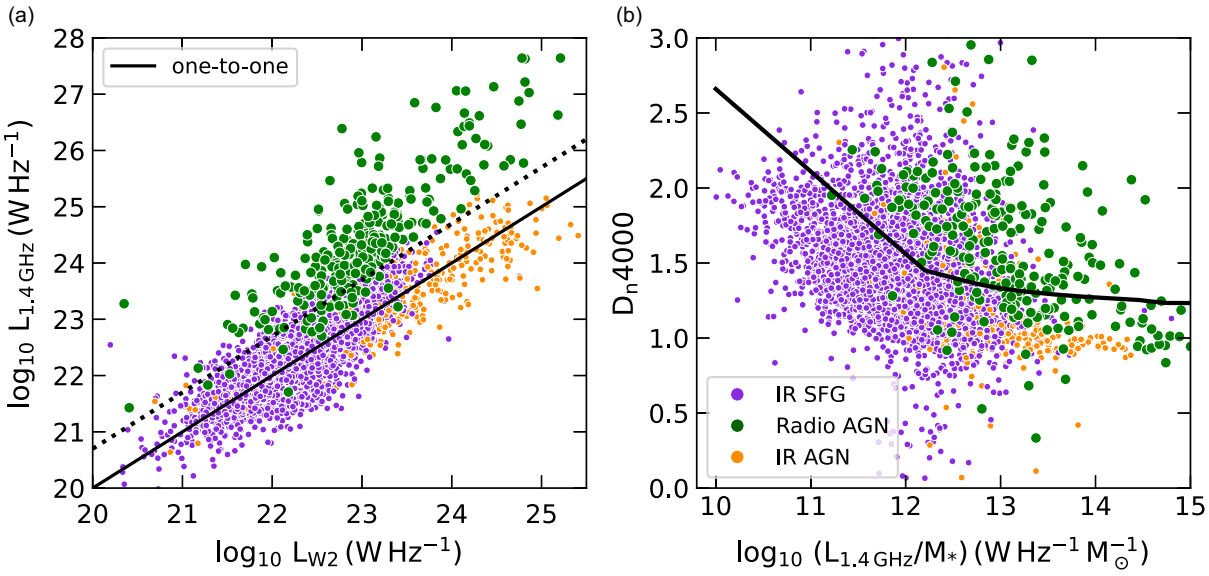


Figure 6. The figure shows the distributions of the IR-radio sources with spectroscopic detections (contains 32% of the IR-radio sample). Panel (a): The IR-radio correlation between the W_2 magnitude and the 1.4 GHz radio flux density. The colours of the dots follow the previous figures. The solid line is the linear relation between the two quantities, and the dotted line corresponds to a version scaled by a factor of 5 (in linear scale). Around 88% of the radio AGN (green dots) lie above the dotted line, implying that the radio emission from a large fraction of the sample is powered by AGN activity. Panel (b): The D_n4000 break index as a function of the 1.4 GHz radio flux density per stellar mass, as used by Best & Heckman (2012). The objects falling below the solid line are most likely low-excitation in nature. It is evident that the radio AGN are scattered around this demarcation line, implying that the optically detected IR-radio sources are of both low- and high-excitation nature.

Table 1. The number of sources in the final samples used in this work. Both optical and IR catalogues are crossmatched with radio sources as described in Section 2.4.

Field	Optical-radio	Infrared-radio
G09	2 091	14 123
G23	2 193	14 339
Total	4 284	28 462

label under the diagnostics (along the rows) include both SFGs and composites and the SFG label under the clustering tools (along the columns) contain both the non AGN clusters.

In the case of SFGs, along the rows in Table 2, it is evident that in most cases ≈ 90% of sources are correctly identified by the clustering tools. Clusters and classes among AGN match at best for ≈ 90% of the sources, with the number dropping to ≈ 40% for the MEx diagnostic (second row of Table 2). To investigate whether this mismatch stems from using $S_{1.4\,\mathrm{GHz}}$ as one of the features, we checked the fractions shown in Table 2 in a parameter space comprised only of optical parameters. No notable difference was detected in the numbers except for small rearrangements in the fractions. This insensitivity to the radio flux density could be interpreted as the mutual exclusion between the astrophysical origins of the optical and radio AGN activity for the sources in our sample. This is no surprise since optical source selection discards a large fraction of the absorption line galaxies, which primarily represent the radio galaxy population, at least in the local Universe (say $z < 0.4$, e.g. Smolčić et al. 2017; Kondapally et al. 2023). This is evident from the number of radio sources with and without optical counterparts used for clustering. From Table 1 and Section 2.2, it can be seen that only ≈ 5% of the total radio sources in both G09 and G23 fields present spectral lines for performing optical diagnostic analysis. This can only be improved using a large scale, deeper spectroscopic survey in conjunction with the EMU data.

6.1.2. *IR and radio parameter spaces*

Table 3 looks at the distribution of sources among the different IR and radio empirical diagnostic tools (see Section 3.2 and 3.3) and the KMeans clustering tool. We do not include other clustering tools here since they do not seem to perform as efficiently as KMeans. We attribute this differing behaviour of the clustering tools to the distribution of the astrophysical classes in the various diagnostic diagrams. The classification schemes in the optical diagnostic diagrams are mostly driven by the density distribution of the data points. Since most clustering tools search for such density distributions, the resulting clusters, for the most part, would agree with the classifications. IR demarcations, on the other hand, are specific cut-offs, which is likely why clustering tools other than KMeans struggle in identifying the astrophysical classification. KMeans does not search for density distributions; rather, it can identify threshold cut-offs as seen in Figure 3(a). It should be noted here that there could be other clustering tools those may have worked as efficiently as KMeans in the scenario presented in this work.

It is evident by now that, in the case of the IR diagnostics, the purple and orange clusters represent the IR SFGs and AGN, respectively. The green clusters, on the other hand, represent the radio AGN as seen in the radio diagnostic (Figure 4). The numbers reported in Table 3 take this into account. The totals in the first three rows do not include the green cluster identified in the respective diagnostics (Figure 3a–3c). The last row of Table 3 includes all three clusters identified in the radio diagnostic (Figure 3d). This is because in the clustering results, there is no KMeans label for the green data points that can be used to put them into either IR SFG

https://doi.org/10.1017/pasa.2026.10175 Published online by Cambridge University Press

Table 2. The distribution of the SFGs and AGN identified by the different optical diagnostic tools among the various clustering tools, with each cell acting as a confusion matrix for a diagnostic-clustering tool pair. The two clusters in the metallicity sequence exhibited by the SFG population (purple and green clusters) have been combined to form a single SFG cluster in the case of the clustering tools. The diagnostic tools are arranged along the rows in the order BPT, MEx, blue, and CEx, and the clustering tools are arranged along the columns in the order KMeans, GMM, FCM, and BIRCH. Approximately 90% of the star forming population identified by the empirical tools are correctly identified by most clustering tools. This fraction comes down to 80% in the case of AGN, with even lower values for the MEx diagnostic (second row, see text for details).

Optical			KMeans (%)		GMM (%)		FCM (%)		BIRCH (%)	
			SFG	AGN	SFG	AGN	SFG	AGN	SFG	AGN
BPT	SFG	3 467	94.89	5.11	91.90	8.10	87.68	12.32	93.37	6.63
	AGN	817	15.91	84.09	21.30	78.70	7.59	92.41	10.41	89.59
MEx	SFG	2 514	98.58	1.42	93.34	6.66	97.15	2.85	97.11	2.89
	AGN	2 106	60.45	39.55	63.01	36.99	46.82	53.18	57.31	42.69
Blue	SFG	4 547	88.15	11.85	85.92	14.08	79.85	20.15	85.92	14.08
	AGN	512	18.56	81.44	23.24	76.76	17.58	82.42	15.82	84.18
CEx	SFG	3 346	96.08	3.92	91.87	8.13	89.51	10.49	93.84	6.16
	AGN	938	21.86	78.14	30.49	69.51	11.41	88.59	19.4	80.60

Table 3. The distribution of the SFGs and AGN identified by the different IR-radio diagnostic tools among the KMeans clusters, with each cell acting as a confusion matrix for a diagnostic-clustering tool pair. It is worth noting the nearly 0% misclassification of the Kozieł-Wierzbowska et al. (2021) radio SFGs as KMeans AGN. The Kozieł-Wierzbowska et al. (2021) radio AGN has a success rate of ≈ 52%, which makes the KMeans clustering selection a stricter, less contaminated AGN selection. The numbers noted in the first three rows (the IR AGN diagnostics) do not include the green cluster population (see text for details).

Infrared-radio			KMeans (%)	
			SFG	AGN
Assef (2018)	SFG	22 400	93.33	6.67
	AGN	4 152	14.38	85.62
Mateos (2012)	SFG	22 767	94.28	5.72
	AGN	3 785	0.98	99.02
Messias (2012)	SFG	25 876	81.71	18.29
	AGN	676	52.96	47.04
Kozieł−Wierzbowska (2021)	SFG	22 925	99.98	0.02
	AGN	5 537	47.51	52.49

or AGN classes. Along the rows, SFGs are identified correctly in at least 90% of the cases, except for the Messias et al. (2012) criterion, where the identification rate is ≈ 80%. This can be attributed to the requirement of K-band magnitudes resulting in a reduced sample size for this diagnostic (see Section 2.4). Also, K-band magnitude was not a clustering parameter, which places a weaker constraint on the cluster identification. These two reasons might be playing the role in driving the difference we see, given that this is a statistical study. AGN classes show varying matching rates, mainly above 80% for the Assef et al. (2018) and Mateos et al. (2012) criteria. For the Messias et al. (2012) diagnostic, where the matching rate reaches ≈47%, a similar explanation as for the SFGs applies here.

It is worth noting the distribution of empirical radio AGN among the KMeans AGN cluster (last row), where the matching rate is close to 52%, which appears relatively low. This, in fact, corresponds to a much stricter radio AGN selection criterion than a simple one-to-one relation prescribed by Kozieł-Wierzbowska et al. (2021) (also see Section 5.2), which is made clear by the negligible (close to 0%) fraction of Kozieł-Wierzbowska et al. (2021) SFGs labelled as KMeans AGN. The remaining ≈ 47% of sources belong to radio star formers (orange and purple clusters) in the KMeans classification, out of which ≈ 56% are IR AGN and ≈ 37% are IR SFGs. A less conservative criterion would redistribute these sources among radio AGN, thereby raising the KMeans radio AGN percentage with respect to Kozieł-Wierzbowska et al. (2021) AGN. This point implies that we can have high confidence in the radio AGN population selected by the IR-radio AGN diagnostics defined in Sections 5.3 and 5.4, since they are defined to select KMeans clusters corresponding to radio AGN.

6.2. A probabilistic view on emission mechanisms

6.2.1. Optical AGN probabilities

Traditional galaxy classification methods work by identifying the dominant emission mechanism, resulting in binary labels for all the galaxies in the sample, a procedure followed throughout this work so far. The simplest, and yet the key point here, is that the contributions from the non-dominant component are still present in the galaxy, and are overlooked. Recent explorations in this direction are provided by Davies et al. (2014a,b), Thomas et al. (2018), and Pennock et al. (2025). Thomas et al. (2018) explored the degree of mixing in the optical spectra of SFGs and AGN using photoionisation models. They found that a median of ≈ 30% of the Balmer line flux arises in star forming like regions, even for galaxies with high values of excitation (log [OIII]/H$\beta \geq 0.9$). Morabito et al. (2025) recently used brightness temperature thresholds to put an upper limit on the radio emission from star formation, thereby isolating regions of radio emission from AGN to derive 144 MHz radio luminosity functions.

In this work, we follow a similar philosophy, where a probabilistic (or fractional) quantification of the different components is preferred over a binary classification of the galaxy energy budget. Some of the clustering tools, such as GMM and FCM, can produce probabilities associated with astrophysical classifications. We infer these probabilities as related to the existence of AGN. As already described in Sections 4.2 and 4.3, GMM considers the data points in the feature space as Gaussian probabilities and assigns the most probable cluster label as a hard classification. FCM follows a similar approach, however, it gives the degree to which a

https://doi.org/10.1017/pasa.2026.10175 Published online by Cambridge University Press

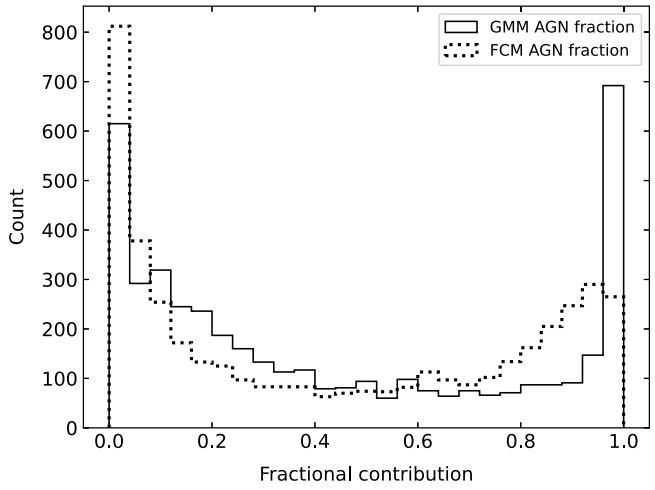


Figure 7. Fractional contributions (or probabilities) of AGN activity and the corresponding number of sources, as reported by GMM and FCM. The solid and dashed lines correspond to the AGN fractions from GMM and FCM, respectively. Since a two cluster approach is adopted, AGN and SFG probabilities in a given clustering tool add up to one. These values can be compared to the relative strengths of the energy generating mechanisms in a given galaxy. As a result, instead of labelling a galaxy either as an SFG or AGN, it can be given a fractional quantification. That is, for instance, 40% SFG and 60% AGN. In a hard classification setting, such a galaxy would most probably be classified as an AGN, still carrying the unaccounted contribution from star formation processes. Methods similar to this work can be used to soft classify galaxies such that the contributions are accounted for without any bias.

data point belongs in a given cluster (in a fuzzy way, see Section 4.3) and assigns the cluster label with the highest degree to that data point. The main difference between these two algorithms is that GMM assumes cluster Gaussianity, whereas FCM relies solely on cluster centroid distances. We choose to keep FCM as well in this discussion since the clusters may not always be Gaussian in reality.

We argued in Section 5.1 that the two clusters lying below the Kewley et al. (2001) line (purple and green) represent star forming species exhibiting the metallicity sequence. We combine the probabilities of these two clusters by performing the clustering for the optical sources using the number of clusters to be two, and present the fractional contribution of star formation and nuclear activity. Two probability values are assigned to each data point by both GMM and FCM, corresponding to their occupancy in one of the two clusters. Figure 7 shows the AGN fractions identified by GMM (solid line) and FCM (dashed line). Both clustering tools recover similar fractions of AGN and SFGs in the optical-radio sample. There is, however, a difference in the number of galaxies identified by both algorithms when it comes to pure SFGs and AGN. This can be attributed to the working principles of GMM and FCM. GMM models the data density, whereas FCM looks at the distances between the cluster centroids and data points. When the clusters strongly overlap, GMM takes the data points in that region to be genuinely belonging to both the clusters, assigning similar probabilities. FCM, on the other hand, looks for the centroid closest to that data point and assigns more weight to that cluster.

Figure 8 shows the distribution of these probabilities in the BPT space, where Figure 8(a) and 8(b) corresponds to the GMM and FCM AGN probabilities, respectively. The colour is mapped to the probability for a given data point to belong to the cluster representing AGN. Since we followed a two-cluster approach, the total probability of AGN and SFGs from a given clustering tool adds up to one. It is evident that AGN exhibit higher probabilities in their respective regions in the BPT diagram. Both GMM and FCM probabilities exhibit strong separation between the SFG and AGN regions, such that the sources with higher excitation values present the maximum probability to be AGN. The probabilities assigned by the clustering techniques look remarkably like a model for the AGN-SFG mixing, similar to the mixing line defined by Thomas et al. (2018), where the data points near the dividing lines have a higher probability to be SFGs and the data points with high excitation values show higher probabilities to be AGN. The similarity between the probability measures from the clustering and astrophysical modelling strengthens the results from clustering, such that these measures can be directly associated with the strengths of the underlying galaxy energy sources.

6.2.2. *IR AGN probabilities*

GMM and FCM are found to produce results that can be compared with the astrophysical classifications only for the optical sources. In the case of IR sources, however, no clustering tool other than KMeans seems to work well, and unfortunately, KMeans does not work probabilistically. As an alternative approach, we looked at the normalised number counts of the different populations identified in the IR-radio space shown in Figure 3(a) as a function of $S_{1.4\,\mathrm{GHz}}$ in one hundred bins, shown in Figure 9. The fractions are calculated as the ratio of the number of sources in a given class and the total number of sources. The green dashed line corresponds to radio AGN (green cluster), the orange dashed line corresponds to IR AGN (orange cluster), and the purple dashed line corresponds to IR SFG (purple cluster). The solid red line corresponds to the total (IR+radio) AGN fraction. These fractions are different from the probabilities discussed in the case of optical sources. GMM and FCM provide the AGN/SFG fractions in a given source, whereas Figure 9 gives the AGN/SFG fraction at a given radio flux density for a sample of sources.

It can be seen that the SFGs dominate the sample at lower flux values ($S_{1.4\,\mathrm{GHz}} < 1$ mJy), where at the lowest fluxes represent $\approx 90\%$ of the radio source population. This fraction of the SFGs at lower radio flux densities compares well with the existing studies (e.g. Seymour et al. 2008; Novak et al. 2017; Pennock et al. 2021; Best et al. 2023). SFGs become the dominant population in the low flux density, most likely because their radio emission is linked to ongoing star formation (e.g. Seymour et al. 2008). The dominance of AGN at high radio flux densities can be attributed to extended radio sources and intrinsically high radio luminosities. The remaining $\approx$ 10% sources at these flux values are IR AGN. Both IR SFGs and AGN start dropping off in numbers above 1 mJy. An increase in the number of radio AGN is also evident above this flux density. These characteristics are in agreement with the observed trends as discussed above. The IR AGN and SFG (possibly radio-quiet) fractions (orange and purple dashed lines) are still expected to carry some radio emission associated with radio activity. As already stated in Section 3.3, even sources without excess radio emission compared to the FIR-radio correlation are found to host radio AGN (see, Panessa et al. 2019). We do not provide probabilities of the IR-radio sources similar to Section 6.2.1. Such an analysis would require assessing catalogue completeness and is aimed at future work.

The philosophy of this paper is encapsulated in this statement, 'every galaxy is a composite.' When galaxies are classified using a binary scheme, the opportunity to explore the complexity of underlying astrophysical processes is often lost. A probabilistic

https://doi.org/10.1017/pasa.2026.10175 Published online by Cambridge University Press

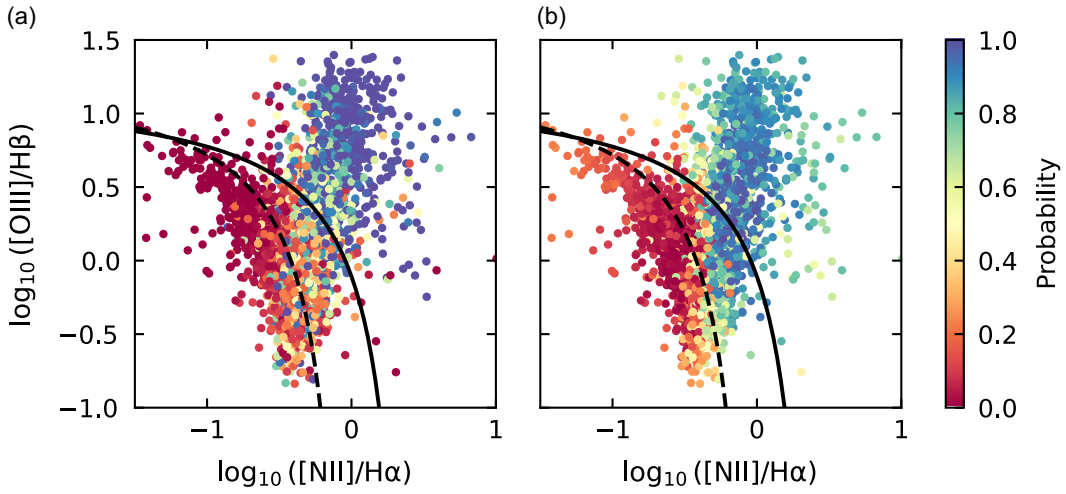


Figure 8. The distribution of the probabilities described in Figure 7 in the BPT parameter space, in the order, GMM AGN in panel a and FCM AGN in panel b. It can be seen that the AGN exhibit higher probabilities in their respective BPT space. The solid and dashed lines correspond to the Kewley et al. (2001) and Kauffmann et al. (2003) lines, respectively.

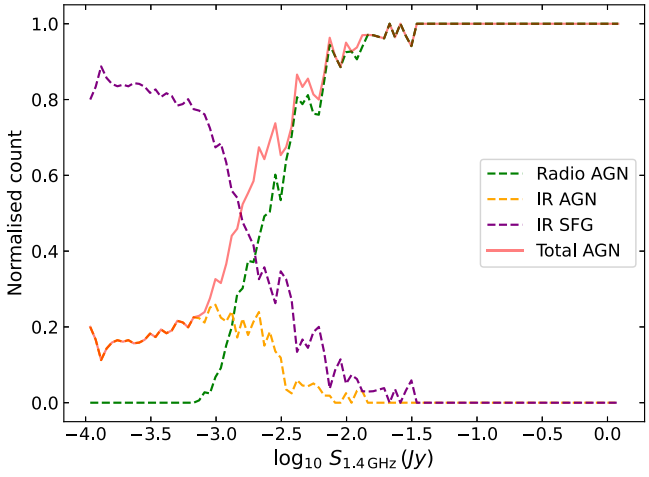


Figure 9. Fractional contribution as a function of radio flux density. The dashed green line represents the radio AGN (green cluster), the dashed orange line represents the IR AGN (orange cluster), the dashed purple line represents the IR SFGs (purple cluster), and the total fraction of AGN at a given radio flux density is represented by the solid red line. It is evident that the star forming fraction is at the maximum below 1 mJy, above which the radio AGN contributes the most to the radio flux budget.

approach, which quantifies the likelihood of a galaxy being star forming or hosting an AGN, allows for a more nuanced interpretation. This enables deeper astrophysical enquiries, such as investigating transitional populations and mixed-mode activity, all of which are obscured by rigid binary classification. In order to emphasise this perspective, we present a catalogue of G09 and G23 radio sources with optical diagnostic labels and probabilities where possible. The catalogue presents the optical-radio sample with GAMA, CatWISE2020, and EMU identifiers, positions, and the probabilities from GMM and FCM. It would also provide the labels given by the four clustering algorithms, where 0 corresponds to SFG and 1 corresponds to AGN. The details of this catalogue are shown in Table 4. These probabilities can be used for deriving luminosity functions encapsulating the exact astrophysics underneath, and can also be used for training ML models, enabling faster and simpler probabilistic classification (see Carvajal et al. in preparation).

7. Conclusions

Galaxies are commonly considered as a population with binary characteristics when discussing their underlying energy generation mechanisms. This is not the whole picture, as evident from observations and theoretical understandings. The main point of this work is to emphasise that the better question to ask is, 'what percentage of a galaxy's emission is powered by its AGN activity?' The benefit is that exploring these fractions will lead to a better quantification of astrophysically relevant quantities (e.g. luminosity functions and densities). The number of AGN identified in binary classifications depends strongly on the wavelength regime and the diagnostic used. This is evident from the different numbers of AGN in the optical and IR samples. This means that the more diagnostics we use to explore the population, the larger the AGN sample size will be, urging a multiwavelength analysis in galaxy population studies.

In this work, we evaluated the performance of different unsupervised clustering tools in the optical and IR-radio diagnostic spaces, using an optical-radio sample and an IR-radio sample, respectively. All four clustering algorithms used, namely KMeans, GMM, FCM, and BIRCH, performed well in the optical diagnostic spaces, whereas only KMeans worked well for the IR-radio diagnostic spaces. We found that the classifications and clusters agree with each other for up to 90% of the time in the case of SFGs and up to 80% of the time in the case of AGN, in both the optical and IR-radio diagnostics. Radio flux density as an additional parameter in the optical diagnostic space did not seem to impact the results. This is probably caused by the sample selection, where the requirement of BPT and other optical lines discards the absorption line galaxies. The IR-radio diagnostics revealed quite interesting results, where one out of the three clusters in the IR space occupied the radio AGN region in the radio diagnostic space.

We used this feature to propose two novel IR-radio AGN diagnostics, (*i*) a two-dimensional space using the $W_1 - W_2$ colour and $S_{1.4\,\mathrm{GHz}}$. This diagnostic selects radio AGN with $\approx$ 90% completeness and $\approx$ 90% reliability, (*ii*) a three-dimensional space that uses the W_2 magnitude as a third dimension to the two-dimensional diagnostic, which also selects radio AGN with $\approx$ 90% reliability and $\approx$ 90% completeness, combining the best of both the

https://doi.org/10.1017/pasa.2026.10175 Published online by Cambridge University Press

Table 4. The AGN probability (or fraction) catalogue columns and their descriptions. The catalogue contains the GAMA, CatWISE2020, and EMU identifiers, positions, EMU flux densities, GMM and FCM probabilities, and the clustering labels, where 0 corresponds to SFG, and 1 corresponds to AGN.

Column name	Unit	Description
component_id	–	Radio source identifier taken from the parent G09 and G23 catalogues
uberID	–	GAMA source id
CATAID	–	GAMA source id (from the GAMA II phase)
CatWISE2020_id	–	CatWISE2020 catalogue id of the radio source
ra_deg_cont, dec_deg_cont	deg	Right ascension and declination from EMU
ra_err, dec_err	arcsec	EMU positional uncertainties
RAcen, Deccen	deg	Right ascension and declination of the optical sources (of the flux-weighted centres, ICRS) from GAMA
RAcen_err, Deccen_err	arcsec	GAMA positional uncertainties
RA_catwise, DEC_catwise	deg	CatWISE right ascension and declination
RA_catwise_err, DEC_catwise_err	arcsec	CatWISE2020 positional uncertainties
Total_flux	milliJansky (mJy)	Integrated radio flux density at 943 MHz
Total_flux_err	milliJansky (mJy)	Uncertainty in the integrated flux density at 943 MHz
prob_agn_gmm	–	Probability for a source to be AGN as per the GMM clustering
prob_agn_fcm	–	Probability for a source to be AGN as per the FCM clustering
KMeans_class	–	KMeans cluster label of the source
GMM_class	–	GMM cluster label of the source
FCM_class	–	FCM cluster label of the source
BIRCH_class	–	BIRCH cluster label of the source

spectral regimes. The superiority of the three-dimensional radio AGN selection is a crucial outcome of this work. Apart from being driven by the mixed parameter spaces of IR and radio, and thereby capturing the essence of both the wavelength regimes, the advent of wider and deeper radio surveys like EMU necessitates the development of new and improved selection criteria for galaxies active in the radio regime. We also publish a catalogue of the G09 and G23 radio sources with the associated probabilities of them being an AGN in the optical regime, emphasising the acute need for considering galaxies to be composed of probabilistic contributions rather than a binary population.

Acknowledgements.. This scientific work uses data obtained from Inyarrimanha Ilgari Bundara/the Murchison Radio-astronomy Observatory. We acknowledge the Wajarri Yamaji People as the Traditional Owners and native title holders of the Observatory site. The Australian SKA Pathfinder is part of the Australia Telescope National Facility (https://ror.org/05qajvd42), which is managed by CSIRO. Operation of ASKAP is funded by the Australian Government with support from the National Collaborative Research Infrastructure Strategy. ASKAP uses the resources of the Pawsey Supercomputing Centre. Establishment of ASKAP, the Murchison Radio-astronomy Observatory and the Pawsey Supercomputing Centre are initiatives of the Australian Government, with support from the Government of Western Australia and the Science and Industry Endowment Fund. This paper includes archived data obtained through the CSIRO ASKAP Science Data Archive, CASDA (http://data.csiro.au).

GAMA is a joint European-Australasian project based around a spectroscopic campaign using the Anglo-Australian Telescope. The GAMA input catalogue is based on data taken from the Sloan Digital Sky Survey and the UKIRT Infrared Deep Sky Survey. Complementary imaging of the GAMA regions is being obtained by a number of independent survey programmes including GALEX MIS, VST KiDS, VISTA VIKING, WISE, Herschel-ATLAS, GMRT and ASKAP, providing UV to radio coverage. GAMA is funded by the STFC (UK), the ARC (Australia), the AAO, and the participating institutions. The GAMA website is http://www.gama-survey.org/.

References

Ahmed, U. T., et al. 2024, PASA, 41, e021
Alonso-Herrero, A., et al. 2006, ApJ, 640, 167
Assef, R. J., et al. 2013, ApJ, 772, 26
Assef, R. J., et al. 2018, ApJS, 234, 23
Baldwin, J. A., Phillips, M. M., & Terlevich, R. 1981, PASP, 93, 5
Baqui, P. O., et al. 2021, A&A, 645, A87
Baron, D. 2019, arXiv e-prints, arXiv: 1904.07248
Belfiore, F., et al. 2016, MNRAS, 461, 3111
Bellstedt, S., et al. 2020, MNRAS, 496, 3235
Best, P. N., et al. 2023, MNRAS, 523, 1729
Best, P. N., & Heckman, T. M. 2012, MNRAS, 421, 1569
Bezdek, J. C., Ehrlich, R., & Full, W. 1984, CG, 10, 191
Birchall, K. L.,Watson, M. G., Aird, J., & Starling, R. L. C. 2023, MNRAS, 523, 4756
Blanton, M. R., & Roweis, S. 2007, AJ, 133, 734
Condon, J. J. 1992, ARA&A, 30, 575
Cowley, M. J., et al. 2016, MNRAS, 457, 629
Cowley, M. J., et al. 2018, MNRAS, 473, 3710
Cutri, R. M., et al. 2012, VizieR Online Data Catalog: WISE All-Sky Data Release (Cutri+ 2012), VizieR On-line Data Catalog: II/311. Originally published in: 2012wise.rept. . ..1C
Cutri, R. M., et al. 2021, VizieROnline Data Catalog: AllWISE Data Release (Cutri+ 2013), VizieROn-line Data Catalog: II/328. Originally published in: IPAC/Caltech (2013)
Davies, R. L., Kewley, L. J., Ho, I. T., & Dopita, M. A. 2014a, MNRAS, 444, 3961
Davies, R. L., Rich, J. A., Kewley, L. J., & Dopita, M. A. 2014b, MNRAS, 439, 3835
de Jong, J. T. A., et al. 2015, A&A, 582, A62

de Souza, R. S., et al. 2017, MNRAS, 472, 2808
Delvecchio, I., et al. 2017, A&A, 602, A3
Delvecchio, I., et al. 2021, A&A, 647, A123
Dempster, A. P., Laird, N. M., & Rubin, D. B. 1977, JRASC, 39, 1
Dias, M. L. D. 2019, Fuzzy-C-Means: An Implementation of Fuzzy CMeans Clustering Algorithm, doi:10.5281/zenodo.3066222
Donley, J. L., et al. 2012, ApJ, 748, 142
Donley, J. L., Rieke, G. H., Pérez-González, P. G., Rigby, J. R., & Alonso-Herrero, A. 2007, ApJ, 660, 167
Driver, S. P., et al. 2011, MNRAS, 413, 971
Driver, S. P., et al. 2022, MNRAS, 513, 439
Eales, S., et al. 2010, PASP, 122, 499
Edge, A., et al. 2013, Msege, 154, 32
Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. 1996, in Second International Conference on Knowledge Discovery and Data Mining (KDD'96). Proceedings of a Conference held August 2–4, ed. D. W. Pfitzner, & J. K. Salmon, 226
Everitt, B. S., Landau, S., Leese, M., & Stahl, D. 2011, Cluster Analysis. Wiley Series in Probability and Statistics. Wiley. https://books.google.com.au/books?id=WSayDAEACAAJ
Fazio, G. G., et al. 2004, ApJS, 154, 10
Fitriana, I. K., & Murayama, T. 2022, PASJ, 74, 689
Fotopoulou, S. 2024, A&C, 48, 100851
Gordon, Y. A., et al. 2017, MNRAS, 465, 2671
Gürkan, G., et al. 2022, MNRAS, 512, 6104
Heckman, T. M., & Best, P. N. 2014, ARA&A, 52, 589
Hermosa Muñoz, L., et al. 2025, arXiv e-prints, arXiv: 2509.14019
Hopkins, A., et al. 2025, PASA, 42, e071
Hotan, A. W., et al. 2021, PASA, 38, e009
Hotelling, H. 1936, Biometrika, 28, 321
Juneau, S., et al. 2013, ApJ, 764, 176
Juneau, S., Dickinson, M., Alexander, D. M., & Salim, S. 2011, ApJ, 736, 104
Karsten, J., et al. 2023, A&A, 675, A159
Kauffmann, G., et al. 2003, MNRAS, 346, 1055
Kellermann, K. I., Sramek, R., Schmidt, M., Sha_er, D. B., & Green, R. 1989, AJ, 98, 1195
Kewley, L. J., Heisler, C. A., Dopita, M. A., & Lumsden, S. 2001, ApJS, 132, 37
Kondapally, R., et al. 2023, MNRAS, 523, 5292
Kormendy, J., & Ho, L. C. 2013, ARA&A, 51, 511
Kozieł-Wierzbowska, D., et al. 2021, ApJ, 910, 64
Lacy, M., et al. 2004, ApJS, 154, 166
Lacy, M., et al. 2007, AJ, 133, 186
Lamareille, F. 2010, A&A, 509, A53
Lamareille, F., Mouhcine, M., Contini, T., Lewis, I., & Maddox, S. 2004, MNRAS, 350, 396
Leahy, D. A., et al. 2019, PASA, 36, e024
Liske, J., et al. 2015, MNRAS, 452, 2087
Logan, C. H. A., & Fotopoulou, S. 2020, A&A, 633, A154
Lyu, J., & Rieke, G. 2022, Universe, 8, 304
MacQueen, J. 1967, in Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability – Vol. 1, ed. L. M. Le Cam, & J. Neyman (Berkeley, CA, USA: University of California Press), 281
Magliocchetti, M., et al. 2017, MNRAS, 464, 3271
Marocco, F., et al. 2021, ApJS, 253, 8
Márquez, I., et al. 2017, FASS, 4, 34
Martin, D. C., et al. 2005, ApJ, 619, L1
Mateos, S., et al. 2012, MNRAS, 426, 3271
Merloni, A., et al. 2024, A&A, 682, A34
Messias, H., Afonso, J., Salvato, M., Mobasher, B., & Hopkins, A. M. 2012, ApJ, 754, 120
Messias, H., Afonso, J. M., Salvato, M., Mobasher, B., & Hopkins, A. M. 2014, A&A, 562, A144
Morabito, L. K., et al. 2025, MNRAS, 536, L32
Nayak, J., Naik, B., & Behera, H. S. 2015, in Computational Intelligence in Data Mining – Volume 2, ed. L. C. Jain, H. S. Behera, J. K. Mandal, & D. P. Mohapatra (New Delhi: Springer India), 133
Norris, R. P., et al. 2011, PASA, 28, 215
Norris, R. P., et al. 2021, PASA, 38, e046
Novak, M., et al. 2017, A&A, 602, A5
Ordovás-Pascual, I., & Sánchez Almeida, J. 2014, A&A, 565, A53
Osterbrock, D. E., & De Robertis, M. M. 1985, PASP, 97, 1129
Panessa, F., et al. 2019, NatAs, 3, 387
Pedregosa, F., et al. 2011, JMLR, 12, 2825
Pennock, C. M., et al. 2021, MNRAS, 506, 3540
Pennock, C. M., et al. 2022, MNRAS, 515, 6046
Pennock, C. M., et al. 2025, MNRAS, 537, 1028
Pilbratt, G. L., et al. 2010, A&A, 518, L1
Prathap, J., et al. 2024, PASA, 41, e016
Prathap, J., et al. 2025, PASA, 42, e077
Rovilos, E., et al. 2014, MNRAS, 438, 494
Sajina, A., Lacy, M., & Scott, D. 2005, ApJ, 621, 256
Sánchez Almeida, J., Aguerri, J. A. L., Muñoz-Tuñón, C., & de Vicente, A. 2010, ApJ, 714, 487
Seymour, N., et al. 2008, MNRAS, 386, 1695
Skrutskie, M. F., et al. 2006, AJ, 131, 1163
Smolčić, V., et al. 2017, A&A, 602, A2
Stern, D., et al. 2005, ApJ, 631, 163
Stern, D., et al. 2012, ApJ, 753, 30
Stevenson, S. D., et al. 2025, arXiv e-prints, arXiv: 2509.06913
Taskesen, E. 2020, pca: A Python Package for Principal Component Analysis
Taylor, E. N., et al. 2011, MNRAS, 418, 1587
Thomas, A. D., et al. 2018, ApJ, 861, L2
Tremonti, C. A., et al. 2004, ApJ, 613, 898
Veilleux, S., & Osterbrock, D. E. 1987, ApJS, 63, 295
Weiner, B. J., et al. 2007, ApJ, 660, L39
Werner, M. W., et al. 2004, ApJS, 154, 1
White, R. L., et al. 2000, ApJS, 126, 133
Wright, E. L., et al. 2010, AJ, 140, 1868
Xu, X.,Wang, J., Li, Z., & Chen, Y. 2024, ApJ, 971, 165
Yan, R., et al. 2011, ApJ, 728, 38
Yang, H., et al. 2022, MNRAS, 517, 5496
Zadeh, L. 1965, IC, 8, 338
Zhang, T., Ramakrishnan, R., & Livny, M. 1996, in ACM SIGMOD Conference
Zhang, Y., & Zhao, Y. 2003, PASP, 115, 1006
Zhang, Y., & Zhao, Y. 2004, A&A, 422, 1113

https://doi.org/10.1017/pasa.2026.10175 Published online by Cambridge University Press