



Universiteit
Leiden
The Netherlands

Machine learning for radio galaxy morphology analysis

Mostert, R.I.J.

Citation

Mostert, R. I. J. (2024, January 25). *Machine learning for radio galaxy morphology analysis*. Retrieved from <https://hdl.handle.net/1887/3715061>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/3715061>

Note: To cite this publication please use the final published version (if applicable).

4 | Finding AGN remnant candidates based on radio morphology with machine learning

Adapted from

Rafaël I.J. Mostert, Raffaella Morganti, Marisa Brienza, Kenneth J. Duncan, Martijn S.S.L. Oei, Huub J.A. Röttgering, Lara Alegre, Martin J. Hardcastle, Nika Jurlin

A&A 674, A208 (2023)

[doi:10.1051/0004-6361/202346035](https://doi.org/10.1051/0004-6361/202346035)

Context. Remnant radio galaxies represent the dying phase of radio-loud active galactic nuclei (AGN). Large samples of remnant radio galaxies are important for quantifying the radio-galaxy life cycle. The remnants of radio-loud AGN can be identified in radio sky surveys based on their spectral index, and identifications can be confirmed through visual inspection based on their radio morphology. However, this latter confirmation process is extremely time-consuming when applied to the new large and sensitive radio surveys.

Aims. Here, we aim to reduce the amount of visual inspection required to find AGN remnants based on their morphology using supervised machine learning trained on an existing sample of remnant candidates.

Methods. For a dataset of 4, 107 radio sources with angular sizes of larger than 60 arcsec from the LOw Frequency ARray (LOFAR) Two-Metre Sky Survey second data release (LoTSS-DR₂), we started with 151 radio sources that were visually classified as ‘AGN remnant candidate’. We derived a wide range of morphological features for all radio sources from their corresponding Stokes-I images: from simple source-catalogue-derived properties to clustered Haralick-features and self-organising-map(SOM)-derived morphological features. We trained a random forest classifier to separate the AGN remnant candidates from the yet-to-be inspected sources.

Results. The SOM-derived features and the total-to-peak flux ratio of a source are shown to have the greatest influence on the classifier. For each source, our classifier outputs a positive prediction, if it believes the source to be a likely AGN remnant candidate, or a negative prediction. The positive predictions of our model include all initially inspected AGN remnant candidates, plus a number of yet-to-be inspected sources. We estimate that $31 \pm 5\%$ of sources with positive predictions from our classifier will be labelled AGN remnant candidates upon visual inspection, while we estimate the upper bound of the 95% confidence interval for AGN remnant candidates in the negative predictions to be 8%. Visual inspection of just the positive predictions reduces the number of radio sources requiring visual inspection by 73%.

Conclusion. This work shows the usefulness of SOM-derived morphological features and source-catalogue-derived properties in capturing the morphology of AGN remnant candidates. The dataset and method outlined in this work bring us closer to the automatic identification of AGN remnant

candidates based on radio morphology alone and the method can be used in similar projects that require automatic morphology-based classification in conjunction with small labelled sample sizes.

4.1 Introduction

Mass-accreting supermassive black holes —also known as active galactic nuclei (AGN)— at the centres of galaxies can form collinear jets, producing synchrotron emission visible at radio frequencies. Such jets are known to impact star-formation rates (SFRs) and therefore play a crucial role in the evolution of their host galaxy.

Through detailed studies of single AGN and particular galaxy clusters, astronomers have identified a number of different stages of radio-loud AGN (see [Morganti 2017](#) for a review): they can be active; dying or remnant; or restarting. The time that AGN typically spend in each of these phases is a key factor in determining how the SFR in a galaxy typically evolves over cosmic time.

Furthermore, determining the ratio of remnant to active AGN will shed light on the rate and magnitude of the adiabatic and radiative energy losses of the radio lobes. With time, the emission of the lobes of an AGN remnant will fall below the detection limit, as a switched-off AGN will not re-energise its lobes with electrons through its jets, while radiative and adiabatic losses continue ([Godfrey et al., 2017](#)).

AGN remnant candidates can be selected by visually inspecting radio galaxies at ~ 100 MHz frequencies. At ~ 100 MHz frequencies, AGN that have recently switched-off should be traceable because the compact features (e.g. the flat spectrum core, jet, or hotspot emission if applicable) have faded while the steep spectrum emission from the radio lobes can remain visible for a longer time. Due to the adiabatic expansion of the lobes, these AGN remnants are expected to have low surface brightness and to be amorphous in shape; bright hotspots in the lobes and jets are expected to disappear ([Morganti, 2017](#)).

Remnants can also be selected based on their spectral index. The synchrotron-based emission from a radio source usually follows a power-law behaviour in its energy spectrum, such that we can describe the relation between the observed flux S and frequency ν as $S \propto \nu^{-\alpha}$, where α is the spectral index. Radio emission from an AGN core is usually flat, while emission from the lobes is often steeper and can become ultra-steep ($\alpha_{150\text{MHz}}^{1.4\text{GHz}} > 1.2$) for AGN remnants.

AGN remnants are relatively rare objects, with estimates suggesting that they make up $< 11\%$ of all radio sources ([Mahony et al., 2016](#); [Brienza et al., 2017](#); [Mahatma et al., 2018](#); [Jurlin et al., 2020](#); [Quici et al., 2021](#)). This means that a lot of radio sources need to be visually inspected to find a remnant, which means that gathering large samples of remnants is very time-consuming. In the era of large radio sky surveys with unprecedented sensitivity, the number of sources to inspect becomes overwhelming.

Until now, the research for AGN remnants has been carried out using individual sources (e.g. [Cordey, 1987](#); [Brienza et al., 2018](#)) or small samples of sources selected based on source-by-source visual inspection (e.g. [Saripalli et al., 2012](#); [Brienza et al., 2017](#); [Mahatma et al., 2018](#); [Jurlin et al., 2020](#); [Quici et al., 2021](#)) or small samples of sources based on spectral criteria (e.g. [Parma et al., 2007](#); [Murgia et al., 2011](#)). Small hand-curated samples of AGN remnant candidates are valuable as they can be complete and clean and therefore justify follow-up observations at higher frequencies to determine whether or not they are actually AGN remnants; see for example the work by [Jurlin et al. \(2020\)](#). However, this time-consuming selection method does not scale to the vast sky areas needed to get large samples that allow stronger constraints on the population properties of AGN remnants and the time an AGN typically spends in the on, off or restarted phase. Large samples of AGN remnant candidates can already be selected based on their spectral index: by creating spectral index maps of multi-frequency observations (e.g. [Harwood et al., 2018](#)), individual (ultra-)steep sources

can be visually inspected (e.g. [Morganti et al., 2021](#)). However, from Monte Carlo simulations, [Godfrey et al. \(2017\)](#) and [Brienza et al. \(2017\)](#) deduce that sources with ultra-steep spectral indices ($\alpha_{150\text{MHz}}^{1.4\text{GHz}} > 1.2$) represent roughly half of the entire AGN remnant population. Selecting all AGN remnants based on spectral criteria alone means that we do not find the other half of the population. Selection based on just spectral criteria would return the entire population only if observations at > 5 GHz —where the spectral steepening occurs sooner— were available for the same large fractions of the sky as our lower frequency data, which is not the case ([Brienza et al., 2017](#)). Finding automated ways to create our sample of AGN remnant candidates based on their morphology would therefore complement automated spectral-index selection and be a necessary step towards a complete census of AGN remnants.

Existing work on the automated sorting of radio galaxies based on morphology can be divided into (discrete) classification systems and (continuous) clustering systems. The classification systems are best used when large numbers (hundreds) of labelled examples are available, for example, to classify Fanaroff–Riley ([Fanaroff & Riley, 1974](#)) class I (FRI) or class II (FR II) galaxies (e.g. [Aniyan & Thorat, 2017](#); [Alhassan et al., 2018](#); [Ma et al., 2019](#); [Mingo et al., 2019](#); [Bowles et al., 2021](#); [Scaife & Porter, 2021](#); [Mohan et al., 2022](#)). Classification becomes more challenging but not impossible when labels are sparse. For example, [Proctor \(2016\)](#) trained oblique decision trees ([Murthy et al., 1993](#)) using features of 48 giant radio galaxies to find more giant radio galaxy (GRG) candidates in the NRAO VLA Sky Survey (NVSS; [Condon et al., 1998](#)). [Proctor \(2016\)](#) compiled a list of the 1, 600 most likely GRG candidates in NVSS. A follow-up by [Dabhade et al. \(2020a\)](#), which included a thorough visual inspection of all the candidates in the list, confirmed 151 GRGs. For sparsely labelled datasets, pre-training on similar labelled datasets (e.g. for optical galaxies: [Walmsley et al., 2022b,a](#)) or semi-supervised learning⁶⁰ can, in theory, be applied (e.g. for radio galaxies: [Slijepcevic et al., 2022](#)). However, designing and implementing a radio galaxy-specific semi-supervised learning setup is not trivial. For example, the setup presented by [Slijepcevic et al. \(2022\)](#) does not yet outperform the simpler, supervised, convolutional neural network baseline by [Tang et al. \(2019\)](#) on real unlabelled radio data. Notable clustering systems based on unsupervised machine learning techniques find similar sources or perform outlier detection using self-organising maps (SOM; [Kohonen, 1989, 2001](#)) or Haralick features ([Haralick et al., 1973](#)). For the application of these techniques to radio galaxies, see [Galvin et al. \(2019\)](#); [Ralph et al. \(2019\)](#); [Galvin et al. \(2020\)](#); [Mostert et al. \(2021\)](#) and [Ntwaetsile & Geach \(2021\)](#), respectively.

In this work, we attempt to reduce the number of radio sources that require visual inspection to create a large sample of AGN remnant candidates based on their radio morphology from the ongoing Low Frequency Array (LOFAR; [van Haarlem et al., 2013](#)) Two-metre Sky Survey (LoTSS; [Shimwell et al., 2017, 2019](#)). Specifically, we focus on sources within the HETDEX area of the sky within the second data release (LoTSS-DR2; [Shimwell et al., 2022](#)), as this area of the sky is accompanied by a high-quality radio source catalogue (see Sect. 4.2).

In the study presented here, we aim to obtain a morphological separation of AGN remnant candidates in the set of well-resolved sources of LoTSS. Visual inspection of all well-resolved radio sources in the HETDEX area of the LoTSS-DR2 catalogue (4, 107 radio sources > 60 arcsec) is extremely time-consuming. Instead, we create a binary machine-learning classifier that suggests the most likely candidates based on a small and incomplete set of already labelled AGN remnant

⁶⁰Semi-supervised learning attempts to learn from a limited set of labels by propagating labels to examples where the model has a high prediction certainty and by generalising the predictive capacity of the model by training on augmented versions of the labelled data.

candidates (Brienza et al. in prep.). By random sampling from our positive and negative model predictions, we estimate the probability of finding more AGN remnant candidates in each set of predictions. If upon visual inspection we succeed in having most AGN remnant candidates in the positive predictions, this means we have created a tool for people to create bigger samples of remnant candidates by visually inspecting only the positive predictions. The methods in this work combine a supervised random forest (RF) classifier — which is a decision-tree-type classifier of the kind used by Proctor (2016)— with features, some of which come from unsupervised learning techniques such as those presented by Mostert et al. (2021) and Ntwaetsile & Geach (2021).

Section 4.2 describes the LOFAR data that we used to find AGN remnant candidates and explains the initial visual inspection. Section 3.4 describes the image pre-processing that we subsequently apply to automatically derive our various morphological features, as well as the setup and training of our RF. In Sect. 4.4, we show the resulting morphological features, the trained RF, and an assessment of the corresponding feature importances. Finally, we discuss the use and limitations of our method in Sect. 4.5 and summarise our conclusions in Sect. 4.6.

4.2 Data

The data we use are confined to an area of 424 square degrees in the HETDEX region of the sky (right ascension 10h45m00s to 15h30m00s and declination 45° to 57°). The observations were taken using the high-band antennas of LOFAR at 120–168 MHz. The LOFAR images with a resolution of 6 arcsec and a median sensitivity of $71 \mu\text{Jy beam}^{-1}$ are part of the LOFAR Two-metre Sky Survey second data release (LoTSS-DR2; Shimwell et al., 2022). The data come with a value-added source catalogue (simply ‘source catalogue’ hereafter) with manual radio source-component associations and a combination of automatic and manual identifications of the corresponding optical host galaxies (Williams et al., 2019) and (Hardcastle et al. in prep.). According to the source catalogue provided by Williams et al. (2019), which is based on the earlier LoTSS-DR1 over the same sky area, the HETDEX region contains 318,520 radio sources.

Visual inspection of radio emission, with the aim being to manually classify a source as an AGN remnant candidate or not, requires judgements as to whether or not a source has a low surface brightness, is amorphous in shape, and lacks signs of a core, jet, or hotspot emission. This is only a realistic possibility for very well-resolved sources, and in this work we therefore only consider the 4,107 radio sources with an apparent angular size of greater than 60 arcsec (ten times the 6 arcsec synthesised beam).⁶¹

At present, the number of AGN remnant candidates in the literature across all radio surveys is below one hundred (e.g. Jurlin et al., 2020; Quici et al., 2021), which is a low number when compared to the hundreds of thousands of detected radio-loud AGN (e.g. Shimwell et al., 2022) and also when compared to the expected fraction of remnants (e.g. Brienza et al., 2017; Godfrey et al., 2017). It is clear that a larger training set of AGN remnant candidates needs to be built. Brienza et al. in prep. applied an automatic rule-based morphological selection, building on the work of Brienza et al. (2017) and Jurlin et al. (2020). After an extensive and time-consuming visual inspection of this

⁶¹The angular sizes, in arcsec, of the radio sources in the public LoTSS-DR1 catalogue are available through the ‘Maj’ or the ‘LGZ_Size’ column. (The ‘Maj’ column indicates that a radio source was created by source finder PyBDSF, while the ‘LGZ_Size’ column indicates that additional crowd-sourced information was used to check if the source finder’s radio component association was correct.) By taking all sources with either ‘Maj’ > 60 arcsec (31 sources) or ‘LGZ_Size’ > 60 arcsec (4,076 sources), we end up with 4,107 sources.

rule-based selection, Brienza et al. in prep. obtained 151 radio sources with > 60 arcsec as likely AGN remnant candidates. Specifically, these labels were based on the morphology of the radio source, but also take into account neighbouring radio emission at 144 MHz (LoTSS Stokes-I). Optical images from the Panoramic Survey Telescope and Rapid Response System 1 (Pan-STARRS1) 3π steradian survey (Kaiser et al., 2010) centred on the location of the radio source were also inspected in order to classify sources with a corresponding nearby star-forming host galaxy as non-candidate. More details about the labelling procedure can be found in the accompanying paper (Brienza et al. in prep.).⁶²

In this manuscript, we take the 151 radio sources labelled as AGN remnant candidates by Brienza et al. in prep., and compare them to the sources with > 60 arcsec in the HETDEX area that have not yet been visually inspected. Until Sect. 4.5, we refer to all these yet-to-be inspected sources (i.e. not including the 151 candidates) as ‘non-candidate’. The different labels are outlined in Fig. 4.1, which gives the respective size and flux distributions of sources in each class.

4.3 Methods

AGN remnant morphological definitions are not absolutely clear and are not readily identifiable based on a single metric. Therefore, we extract and test a wide range of morphological features in a reproducible manner: namely simple source properties, morphology metrics normally used in optical astronomy, Haralick features, and SOM-derived features (Alegre et al., 2022). The SOM-derived features are created by building on the rotation and flipping invariant SOM algorithm implemented by Polsterer et al. (2015) and built upon by Mostert et al. (2022). In this work, we add source-size invariance to the rotation and flipping invariant SOM. We also introduce a ‘compressed-SOM’ technique that allows better generalisation of the trained neurons. We then train a RF to use the morphological features to classify radio sources as AGN remnant candidate or non-candidate. To establish how well the classifier is able to leverage the features to predict these labels, we use label permutations. Furthermore, we quantify which morphological features are most influential in detecting AGN remnant candidates using feature permutations.

Our hypothesis is that the positive predictions from the trained RF are more likely to be labelled ‘AGN remnant candidate’ upon visual inspection than the negative predictions. This is irrespective of what a source turns out to be after visual inspection; a ‘positive’ prediction simply means that the model predicts the label ‘AGN remnant candidate’ and a ‘negative’ prediction means that the model predicts the label ‘non-candidate’. By visually inspecting random samples of positive and negative predictions, we can test this hypothesis. If the hypothesis is true, we will be able to reduce the number of radio sources that require visual inspection when creating a large sample of AGN remnant candidates based on morphology by only having to inspect the positive predictions of our model. An overview of our approach is visualised in Fig. 4.2.

4.3.1 Data pre-processing

The first step in our automated AGN-candidate-suggestion method is to reduce the observed Stokes-I image maps to a single, pre-processed image cutout per radio source. These cutouts can

⁶²All data, including the list of 151 AGN remnant candidates are at https://lofar-surveys.org/finding_agn_remnants.html.

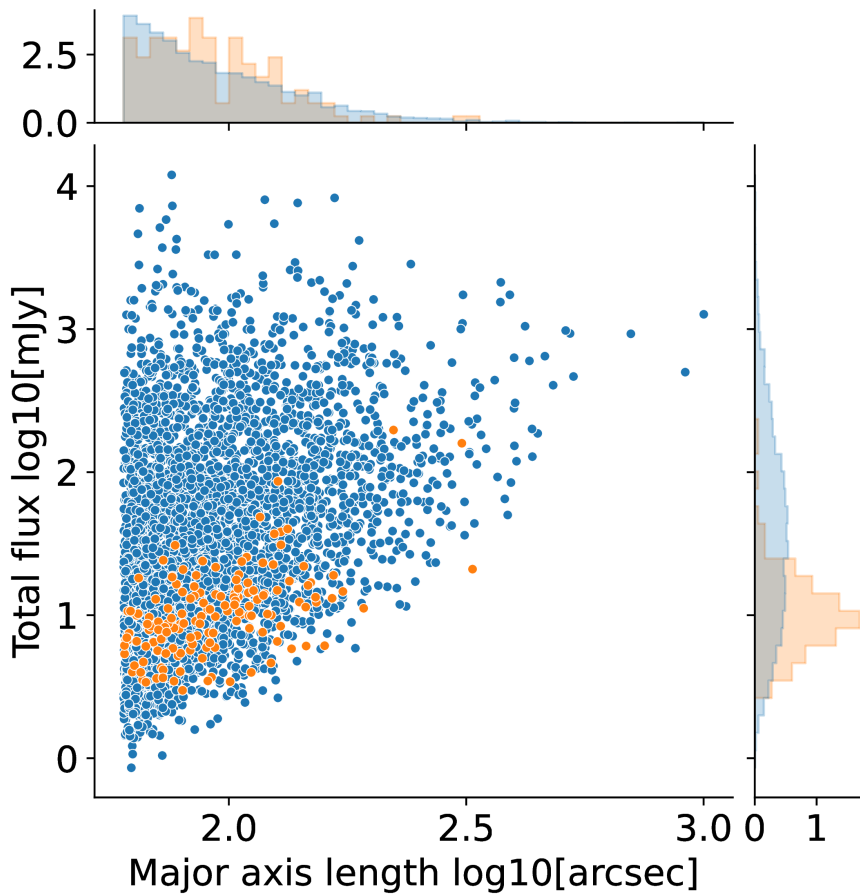


Figure 4.1: Source size versus total flux density for the sources in HETDEX, > 60 arcsec labelled as ‘AGN remnant candidate’ following initial visual inspection by Brienza et al. in prep. (orange dots) and ‘non-candidates’ (blue dots).

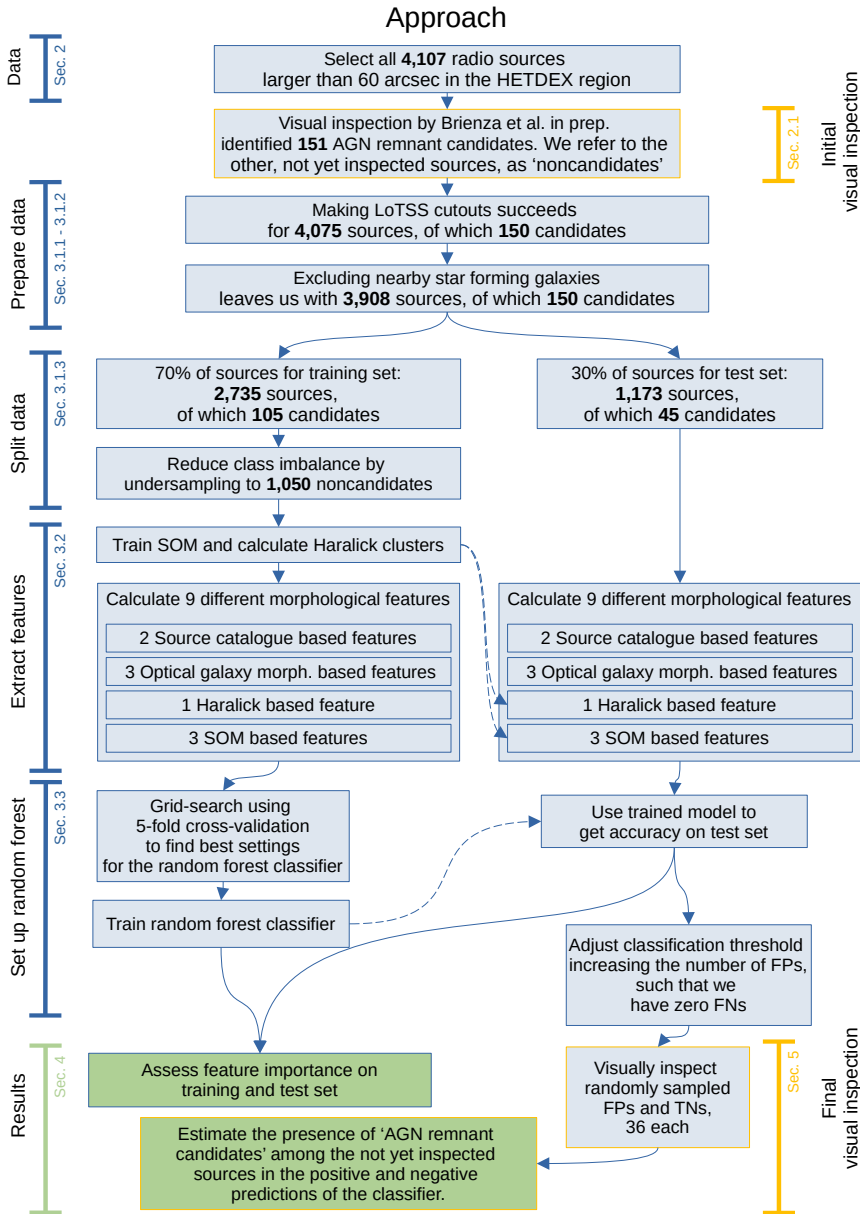


Figure 4.2: Diagram of our approach to create a model that classifies radio sources as either ‘AGN remnant candidate’ or ‘non-candidate’ based on a number of morphological features. The solid arrows show the flow of our data, while the dashed arrows indicate the use of various trained models, specifically a SOM, HDBSCAN* clustered Haralick features, and a RF, all trained using only the data from the (undersampled) training set. The brackets indicate the different stages of our method and mention the sections in which corresponding details can be found.

later be used to automatically extract morphological features for each radio source. The second step in our methods is to automatically exclude most nearby star-forming radio galaxies (SFGs) for which the radio emission is associated with star formation rather than (past) AGN activity.

Creating image cutouts

Apart from the source-catalogue-based features, all morphological features are based on the image of the considered radio source. We therefore extracted image cutouts for each of the 4, 107 sources. Specifically, we created square cutouts from the Stokes-I LoTSS-DR2 image pointings centred on the corresponding sky coordinates in the source catalogue with a size of 1.5 times the source size as reported by the source catalogue. These variable-sized images were then all resized to a fixed size of 174×174 arcsec² images using bilinear interpolation. This ensures that the various morphological features derived downstream are always comparing sources with roughly the same extent in pixels, thereby approximating source-size invariance. As the angular pixel resolution of the LoTSS images is 1.5 arcsec per pixel, this results in images of 116×116 pixels². To prevent radio emission in the corner of these square cutouts from influencing the morphology measures, we mask all pixels outside the circular aperture with a diameter of 116 pixels centred on the middle of the cutout. The cutout-extraction process fails for sources for which the Stokes-I image contains NaNs. The 32 sources for which cutout-extraction failed will be discarded from our dataset. Thus, after cutout extraction, we are left with 4, 075 sources of which 150 AGN remnant candidates.

Mostert et al. (2021) reduced the effect of noise in the radio source image cutouts by clipping away signal that is less than 1.5 times the local noise. In this work, we do not apply such a clipping because AGN remnants are likely to have a surface brightness close to the noise. To prevent the morphological features that we derive later on from providing different outcomes for sources with similar morphologies but different apparent brightness, we rescale all pixel values p_i in all images to $p_{i,\text{norm}}$ according to Eq. 4.1:

$$p_{i,\text{norm}} = \frac{p_i - \min(p_i)}{\max(p_i)}. \quad (4.1)$$

Unlike the work presented by Mostert et al. (2021), we use a radio catalogue that includes manual and likelihood ratio-based source-component association (Williams et al., 2019) as the starting point for creating cutouts, and therefore assume that the source-component association is mostly correct. This means that we are less prone to mistaking a spatially separated single radio lobe for an amorphous AGN remnant candidate. Moreover, this allowed us to remove neighbouring unrelated radio components that fall within our square image cutouts. We did so by subtracting the Gaussian components that, according to the Python Blob Detection and Source Finder (*PyBDSF*; Mohan & Rafferty, 2015) source model, constitute the neighbouring sources from the Stokes-I image cutout.

Exclusion of nearby star-forming radio galaxies

A single amorphous radio blob at 144 MHz might indicate the presence of an AGN remnant, but the same morphology can also signify synchrotron radiation from the supernovae in the star-forming regions from a nearby galaxy without any (past) AGN activity; see Fig. 4.3 for an example. The extent of the radio emission we observe must be far larger than the extent of the optical or infrared emission to ensure that the radio emission originates from (past) AGN activity, and not from a nearby SFG. After trial and error, we adopted the criterion that the LoTSS source major axis must

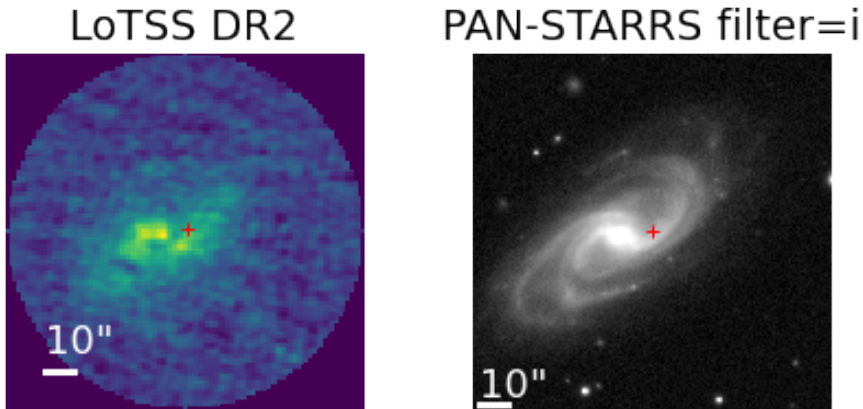


Figure 4.3: Radio source that can be rejected as a AGN remnant candidate as the optical image, taken by Pan-STARRS_i, reveals that the corresponding optical host of the radio source is a nearby star-forming spiral galaxy.

be at least ten times larger than the optical size given by the extended source catalogue (2MASX; Jarrett et al., 2000) from the Two Micron All Sky Survey (2MASS; Skrutskie et al., 2006). From the 4, 075 successfully extracted sources > 60 arcsec, 167 sources were labelled as potential nearby SFGs using this method. After visual inspection, comparing the LoTSS and the Pan-STARRS_i images, we find that this label is correct for 86.2% (144/167) of the sources.

When classifying radio sources as AGN remnant candidates in even larger datasets in the future, we will not visually inspect these potential nearby SFGs but simply discard them. In this work too, we discard all 167 sources labelled as potential nearby SFGs, leaving us with 3, 908 radio sources of which 150 candidates. This means that we do not consider the $(167 - 144)/4075 = 0.6\%$ of sources > 60 arcsec that could have been AGN remnant candidates.

An alternative method to exclude nearby SFGs is to use the optical and near-IR information to derive SFRs and to compare the corresponding expected radio continuum flux density through the radio–SFR correlation (Gürkan et al., 2018) with the observed flux density. However, we decided not to adopt this method as it is only reliable for sources with sufficient multi-wavelength information (e.g. Smith et al., 2021).

Splitting data into a training and a test set

Statistics derived from applying a model to the same data that was used to train the model are potentially unrealistically optimistic due to (possible) overfitting. We therefore split 70% of our data into a training set and used the remaining 30% of the data for our test set. We used a stratified split, which means we randomly assign each radio source to either the training or test set while roughly maintaining the same class balance (‘candidate’ or ‘non-candidate’) as in the full dataset, leaving us with a training set that contains 2, 630 ‘AGN remnant candidates’ and 105 ‘non-candidates’.

Like many classifiers, an RF achieves a poorer performance on an imbalanced dataset. An imbalanced dataset is a dataset for which the number of examples is not evenly spread over the different classes. In our case, for each AGN remnant candidate in the data that we work with, there are more than 26 sources labelled ‘non-candidate’, thereby making it easier for a classifier to achieve high ac-

curacy by always predicting that a given radio source is a ‘non-candidate’. There are broadly three ways to minimise the effect of an imbalanced dataset. One can oversample the minority class, undersample the majority class, or adjust the weight of each sample during training based on its class (Murphy, 2012; Ivezić et al., 2019). Which of these approaches works best depends on the dataset used. We combined undersampling the majority class with class-weight adjustment (for details on the latter, see Sect. 4.3.3). Specifically, we undersampled the majority class in the training set from 2,630 to 1,050 sources, increasing the ratio of AGN remnant candidates to non-candidates in the training set from 1 in 26, to 1 in 10, thereby helping our classifier to learn the features of the less numerous candidate. Hereafter, ‘training set’ refers to the 1,155 sources in our undersampled training dataset and the ‘test set’ to all 1,173 sources in our test dataset.

4.3.2 Deriving morphological features

In this section, we detail how we automatically derive nine morphological features for the sources in the training and test set. These nine features will feed into our classifier on a per-source basis in order that it may predict whether or not the source is a likely AGN remnant candidate.

Source-catalogue-derived properties

From the radio source catalogue produced by Williams et al. (2019), we first calculated the ratio of the length of the major axis to the length of the minor axis of the source (as projected on the sky plane), as we expect that AGN remnants to have a lower ratio due to adiabatic expansion of their lobes in all directions combined with an absence of a longitudinally outward-pushing jet. Second, we calculated the ratio of the total flux density to the peak flux density of the sources, as we expect AGN remnants to have a low and uniform surface brightness due to the adiabatic expansion and the absence of jet-formed hotspots in the lobes. We do not use the total flux density from the catalogue as a feature, because this metric does not inform us about intrinsic brightness and will bias our classifier towards sources with a low apparent flux.

Applying optical galaxy morphology metrics to radio sources

We included three morphological metrics that are normally used in quantifying optical galaxy morphology to see whether they are able to help separate AGN remnant candidates from the not yet inspected sources when applied to radio morphology. Specifically, we implemented the concentration index, clumpiness index, and Gini coefficient as described by Conelice (2014).

The concentration index (C) is defined as the radius at which 80% of the flux of a source is enclosed (r_{80}), divided by the radius at which 20% of the flux is enclosed (r_{20}):

$$C := 5 \log \left(\frac{r_{80}}{r_{20}} \right). \quad (4.2)$$

As expected, we see (Appendix 4.A) that low concentrations match with sources that have little flux in their geometrical centres, meaning they are either FRIIs or strongly bent FRIs. High concentration indices match with sources that have more flux in their geometrical centres, meaning FRIs and diffuse objects. Using the concentration index, a qualitative inspection shows that it (reassuringly)

separates sources where the brightness is core-dominant from those where the brightness is lobe-dominant.

The clumpiness index (S) is defined as:

$$S := 10 \frac{\sum(I - I^\sigma)}{\sum I}, \quad (4.3)$$

where we have a pixel-wise subtraction of the radio source image I and the same image smoothed by a Gaussian kernel I^σ with a standard deviation that is 12 arcsec (two times the synthesized beam). Sources with smooth surface brightness should have low clumpiness values, while sources with small-scale structures should have high clumpiness values. The smaller the angular extent of the sources, the less structure they can exhibit (we should not be able to discern features smaller than our 6 arcsec beam). We expect AGN remnant candidates to have a smooth brightness distribution due to the lack of compact structures, leading to low clumpiness values. We also expect our remnant candidates to be faint, as the particles in the lobes radiate their energy away while they are, by definition, not replenished in an AGN remnant. However, for sources that barely surpass the noise level, the noise pattern might be picked up by the clumpiness index, leading to higher clumpiness index values.

To determine the Gini coefficient of an image, we sort the pixel values (f) in ascending order within a circular aperture that contains 95% of the total flux density within the cutout. The Gini coefficient (G) is then given by

$$G := \frac{1}{|\bar{f}|n(n-1)} \sum_i^n (2i - n - 1)|f_i|, \quad (4.4)$$

where \bar{f} is the average pixel value within the circular aperture of the image. The Gini coefficient is 1 if all flux is concentrated in one pixel, and zero if every pixel contains an equal amount of flux. The Gini coefficient behaves similarly to the clumpiness index: sources with highly concentrated surface-brightness areas, such as sources with bright core or lobe emission, will receive high Gini coefficient values, while diffuse sources and sources close to the noise level will receive low Gini coefficient values (see Appendix 4.A).

Haralick-based morphological features

Haralick texture features, Haralick coefficients, or simply Haralick features are statistics that describe an image or image patch based on the spatial correlation of different pixel intensities within the image (Haralick et al., 1973). Ntwaetsile & Geach (2021) used Haralick features to cluster radio sources and detected outlying sources based on their radio morphologies, motivating us to include Haralick-based morphological features in our classifier. Haralick et al. (1973) define 14 statistics, the first 13 of which are computationally stable, and so we only use those, all of which are based on a grey-level co-occurrence matrix (GLCM). The GLCM is a square matrix with a width and length equal to the number of discrete grey levels in the image (256 in our case). Each value $p_{i,j}$ in this matrix, with i the row index and j the column index, is the probability of a transition from a pixel with a grey-value i to a pixel with a grey-value j in the direction θ at a distance of d pixels. We refer the reader to Haralick et al. (1973) for the 13 statistics subsequently extracted from the GLCM.

We calculated the Haralick features using the Python implementation by Coelho (2013).⁶³ The images we use are the pre-processed Stokes-I cutouts described in Sect. 4.3.1, with the additional step that we reduce each pixel value to 8-bit as in the work of Ntwaetsile & Geach (2021), resulting in pixel values ranging from 0 to 256 in integer steps only, as a meaningful GLCM can only be constructed using integer values. Following Ntwaetsile & Geach (2021), we adopt a distance $d = 1$ and use the offset directions up, down, right, and left for θ , thereby ensuring that the subsequently derived statistics are at least equivalent for any integer number of 90 degree rotations of the input image. We thus ended up with Haralick features: 13 continuous values per source.

If more data labelled as AGN remnants were available, we could incorporate all 13 Haralick features. However, with our current sample, adding all of them would massively enlarge the feature-space, leading to overfitting. Therefore, instead, we proceeded to cluster the Haralick features akin to the method laid out by Ntwaetsile & Geach (2021) using a HDBSCAN* algorithm.

HDBSCAN* (Campello et al., 2015) is a clustering method based on density; it groups data points that have many nearby neighbours to a cluster. The number of clusters is not preset as with a k-nearest neighbours algorithm but is an emergent property based on the data and the chosen hyperparameters. Apart from these (high-density) clusters, HDBSCAN* also creates a single outlier or ‘noise’ cluster to which data points without many nearby neighbours are all assigned.

Specifically, we used the Python HDBSCAN* implementation created by McLInnes et al. (2017). We adopted the HDBSCAN* implementation default parameters except for the ‘minimum cluster size’ and the ‘minimum sample size’, which, together with the data input dimensions, influences how many clusters are formed. According to Ntwaetsile & Geach (2021), the desired number of clusters as influenced by these parameters is somewhat arbitrary and can be decided by the user; the authors suggest setting both values to 64. With our data, this leads to only two clusters apart from the noise cluster; we therefore decided to halve both parameters to 32, leading to the formation of six clusters plus the noise cluster.

The feature that we fed into our classifier is the ratio of AGN remnant candidates in the Haralick cluster of a particular source. It is important to note that this feature changes as a function of the dataset for which it is evaluated. Calculating this feature in advance would therefore lead to data leakage, which leads to over-optimistic results and reproducibility issues (Kapoor & Narayanan, 2022). Therefore, we adapted the classifier training, validation, and inference such that this feature is calculated using only the data available within the specific training, validation, or test sets that we employ.

SOM-derived morphological features

A SOM is a single-layer neural network, made up of neurons arranged on a lattice. It has the property that, during training, each neuron will become increasingly similar to a subset of the training inputs. After training, each neuron can therefore be said to represent a certain set of inputs. Within a well-trained SOM, neurons with similar properties will be located close to each other on the lattice (Kohonen, 2001).

After training, each input from the training set can be ‘mapped to the SOM’: it can be assigned to the neuron to which it is most similar. The same can be done with images outside of the training set, as long as they go through the same preprocessing steps and are not outside of the distribution. We can, for example, take images from the same survey, but in a different area of the sky and

⁶³<http://pypi.python.org/pypi/mahotas>

Table 4.1: Hyperparameters used for training the SOM

SOM lattice dimensions (w x h x d)	$9 \times 9 \times 1$
Number of channels or layers	1
Training image dimensions	116×116 pixels ² , or 82×82 pixels ² after rotation
Neuron image dimensions	82×82 pixels ²
Neighbourhood radius start θ_0 decrease θ_d	5 0.9
Learning rate start α_0 decrease α_d end α_e	1 0.8 0.3
Periodic boundary conditions	False
Number of training epochs	25
Initialisation	Zeros

map these to the trained SOM. By looking at the number (and ratio) of AGN remnant candidates mapped to each neuron of a trained SOM, we can find to what extent which neurons are more indicative of remnant-morphology than others. If the absolute number, or the ratio of remnants is high in a particular neuron, than that neuron is more indicative of remnant-morphology.

Training the self-organising map To use a SOM, we must first train the SOM using a set of inputs from our training set. In our case, the inputs are the Stokes-I image cutouts centred on the 1, 155 radio sources in our training set (see Sect. 4.3.3) for which we detailed the selection and preprocessing in Sect. 4.2. This means that the SOM does not use the source catalogue, but only takes the pixel information of each source into account. Training establishes the properties of a SOM: each neuron will come to represent a subset of images from the training set, and similar neurons are located close to each other on the lattice. At each step in the SOM training procedure, a training image is compared to all the neurons in the map using the Euclidean norm between the training image and the neuron. The pixel values of the neuron with the lowest Euclidean norm to the image are then slightly (controllable with the ‘learning rate’ hyperparameter α) updated to match the training image. The neighbouring neurons are also updated to match the training image by an amount that decreases with increasing distance from the best-matching neuron (controllable with the ‘neighbourhood radius’ hyperparameter θ). The procedure we used to train the SOM is to a large extent analogous to the procedure detailed by Mostert et al. (2021); see Table 4.1 for the hyperparameter values we used for training.

The original SOM training and mapping algorithms are not equivariant under rotation or flipping, while this equivariance is crucial for clustering radio galaxy morphologies. Polsterer et al. (2015) achieve rotational and flipping invariance with respect to the training images by comparing each neuron with many rotated and flipped copies of the training image. We used the Parallelized rotation and flipping INvariant Kohonen maps (PINK) code developed by Polsterer et al. (2015) and introduced an approximate form of angular size invariance by resizing the central source in each image to the same fixed size using bilinear interpolation (as explained in Sect. 4.3.1), before feeding the image to the SOM. We discuss the effect of resizing sources on the SOM in Sect. 4.5.

The size of a SOM lattice is a user-set hyperparameter, and is arbitrary to a certain extent. For our SOM, we adopted a 2D lattice consisting of $N = 9 \times 9$ neurons. To allow neurons to emerge that represent a wide range of common morphologies from the training images, one can set the neighbourhood radius to be very small, such that updates mostly affect a single neuron. However, it is easy to lose coherence in this way, which is to say that multiple (distant) neurons may independently

start to represent the same common morphology. A larger lattice (a lattice with more neurons) with a higher neighbourhood radius preserves the coherence and allows a wide range of morphologies to emerge in the neurons. However, this will also increase the number of redundant neurons, that is, neurons that are close by and roughly represent the same set of sources. Additionally, if the number of sources in our calibration set is of the same magnitude as N , the best-matching neurons to the sources will not be very robust: minor morphological changes to an input might cause the source to be mapped to a neighbouring neuron. This is likely to negatively affect the informativeness of the morphological features that we derive from our SOM. These features are the absolute and relative number of calibration sources that are also mapped to the same best-matching neuron of a particular source.

As a compromise, we trained our SOMs using a 9×9 lattice, preserving coherence and allowing a wide range of common morphologies to be represented, after which we prune the SOM by removing every other row and column. This leaves us with a 5×5 lattice; each of the neurons in this ‘compressed’ SOM is relatively more unique than the ones in the 9×9 SOM. Moreover, the smaller lattice allows the morphological features that we derive from the SOM to be more robust as each neuron contains more sources.

Mapping to the compressed SOM After training, we mapped all sources to the trained, compressed, 5×5 SOM. Here, by ‘mapping’, we mean that we compare an input to each of the trained neurons using the Euclidean norm, and identify the lattice coordinates of the neuron to which it has the smallest Euclidean norm. The features that we used in our RF are the absolute number of AGN remnant candidates mapped to the same neuron best-matching the considered source, and the ratio of AGN remnant candidates to all sources mapped to that neuron.

Similar to the Haralick cluster ratios, the number of AGN remnant candidates per mapped neuron and the ratio features change as a function of the dataset for which these features are evaluated. Here, again, we adapted our classifier training, validation, and inference to prevent data leakage.

Using the SOM similarity metric as a morphological feature We used the Euclidean norm as the similarity metric to decide which SOM neuron best represents a certain input, both during the SOM training and during the SOM mapping stage. The Euclidean norm leads to a continuous value that is larger if the difference between the pixels of the input and the pixels of the neurons increases. By assigning an input to its best-matching neuron, this Euclidean norm is reduced from a continuous value to the discrete coordinate of the neuron on the SOM lattice. However, the continuous value could still provide us with more information. Even within the set of sources mapped to a single neuron, there are differences in the Euclidean norms (in other words, there are differences in how well they morphologically match this neuron). If the Euclidean norms of the calibration inputs lie in a specific range, the RF could use this to classify our sources as AGN remnant candidates. We therefore extracted the Euclidean norms of all our sources as a feature that the RF can use.

4.3.3 Training the random forest classifier

After deriving the different morphological features, we wanted to test the relationship between these features and our initial labels. Specifically, we wanted to train a binary classifier that uses our derived features to predict whether or not a radio source is likely to be an AGN remnant candidate; we wanted to test the predictive power of this classifier (how much better than chance is it), we

wanted to see how it performs on unseen data, and we wanted to establish which features influence this classifier the most.

A decision tree is one of the few classifiers that can handle both discrete feature inputs (like the absolute number of AGN remnant candidates for a specific neuron) and continuous feature inputs (such as the total-to-peak flux ratio of a source). Furthermore, a decision tree is robust to outlying feature-input values, and is insensitive to monotone transformations of its feature inputs, thus allowing our features to be vastly different in terms of absolute magnitude and value range (Murphy, 2012; Ivezić et al., 2019). The disadvantage of a decision tree is that it has a high variance, as features that are high up in the tree have a disproportionate effect on its outcome. This disadvantage can be reduced by using an ensemble of decision trees. An RF classifier returns the average over the outcome of many decision trees, each of which randomly samples only a fraction of the available features, thereby reducing the variance introduced by a single decision tree (Breiman, 2001). For these reasons, we choose to use an RF classifier; specifically, we use the RF as implemented by the sci-kit learn Python library (Pedregosa et al., 2011a).

An RF has a number of hyperparameters to tune; these are user-set parameters that impact the performance of a classifier but are not directly learnt when training a classifier (or any other machine-learning estimator). The optimal hyperparameter values depend on the dataset used and the metric that the user wishes to optimise for. Optimal hyperparameter values can therefore be approximated by testing a range of reasonable values for the desired metric on an independent subset of the data (the validation set). The most important hyperparameter for an RF is the number of trees in the forest, where a larger number leads to longer computation times but also leads to lower variance up to a certain point. Using 5,000 trees allows us to tune the other hyperparameters of the tree in a time on the order of minutes. With our relatively small number of data samples (radio sources) and many features, we are prone to overfitting our data. To reduce the tendency to overfit, the maximum depth of each tree in the RF can be limited and the maximum number of features to use per tree can be altered (Breiman, 2001). (Notably, RF does not overfit as a function of the number of trees (Breiman, 2001).) To further address the class imbalance in our dataset, we can vary the weight of each sample with respect to the weight of samples from the majority class. We employed a grid search on our training set to find good values for these hyperparameters. Specifically, we checked the following maximum depth values: [4, 8, 16, 32, 64, 128, 256, 512, 1024], the following maximum feature-ratios [0.2, 0.3, 0.4, 0.5, 0.6], and the following class weight ratios [0.01, 0.04, 0.16, 0.32, 0.5].⁶⁴ We set the grid search to optimise for F_2 -score, as finding most AGN remnant candidates in large samples is our main objective. An F_1 -score strives for a balance between precision and recall, but as we are more interested in recall than precision, we used the F_2 -score, which is a special case of the F_β -score (with $\beta = 2$) defined as:

$$F_\beta = (1 + \beta^2) \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}. \quad (4.5)$$

We find the best hyperparameter values when using a five-fold stratified cross-validation on the training set, where ‘stratified’ means that the class balance in each cross-validation fold is similar to the class balance in the training set. For all other hyperparameters, we adopted the default values of the

⁶⁴A class weight ratio x indicates that samples from the majority class are weighted by x , while samples from the minority class are weighted by $1 - x$. The low values of x that we explore in our grid search imply that we down-weight the examples from the majority class to increase the exposure of the minority class (for which we have fewer samples) to the classifier.

Pedregosa et al. (2011a) RF implementation.⁶⁵ Finally, we trained an RF with these optimal hyperparameter values using all 1, 155 sources in the training set (as opposed to using one-fifth of the training sources as we did during cross-validation).

4.4 Results

Following feature extraction and the RF setup, we can first explore the position of the AGN remnant sample within our feature space. We can then look at the RF training results.

4.4.1 Resulting morphological features

Figure 4.4 shows the distribution of values we obtained for the Euclidean norm, Gini coefficient, clumpiness index, and concentration index for sources in the training set. Generally, the AGN remnant candidates (orange) overlap with the noncandidates. However, the AGN remnant candidates seem to have higher Euclidean norm values, and lower Gini coefficients, clumpiness indices, and concentration indices. The high Euclidean norm values are not unexpected, because AGN remnants are rare and are therefore not well represented by the SOM (which leads to high Euclidean norm values). The lower Gini-coefficients and clumpiness indices are expected as remnants have a smoother surface brightness than regular AGN. The lower clumpiness indices can be expected due to the absence of significant core-emission and hotspots for AGN remnants. The figures in Appendices 4.A and 4.B show examples of radio-source cutouts in different ranges of Gini coefficient, clumpiness index, concentration index, and different Haralick clusters. The figures show that the features are able to separate sources with different morphological properties, but they also show that the surrounding noise(-pattern) and residual emission from (artefacts of) neighbouring sources have a strong impact on the feature values, making classification based on these features harder.

4.4.2 Radio morphology in the trained SOM

Figure 4.5 shows the results from our SOM-derived features. By inspecting the SOM —the left-most panel of Fig. 4.5— we observe a key property of SOMs: neurons of the SOM that are close together on the 9×9 map are also similar in the morphologies that they represent and the transition between different morphologies is smooth. As a result, we can see smooth gradients of changes in morphologies. The neurons close to the top-left to bottom-right diagonal show morphologies representing FRI-like (core-dominated) sources, while away from this diagonal we see neurons representing FRII-like sources with edge-brightened lobes. Going from the top left to the bottom right, we see neurons representing morphologies that gradually go from less to more elongated (i.e. the neurons represent sources going from a higher to a lower width-to-length ratio). Additionally, we note that in the bottom right, the neurons represent sources that are far above the local noise, while in the top left, the neurons represent sources with surface-brightness levels closer to the local noise.

⁶⁵Tuning the maximum size of decision trees in the RF is a way to trade-off performance versus overfitting. We arbitrarily chose to do so through the maximum RF depth, but a similar effect can (alternatively or additionally) be achieved by tuning the maximum number of leaf nodes, or tuning the minimum number of samples required for a node split in the RF.

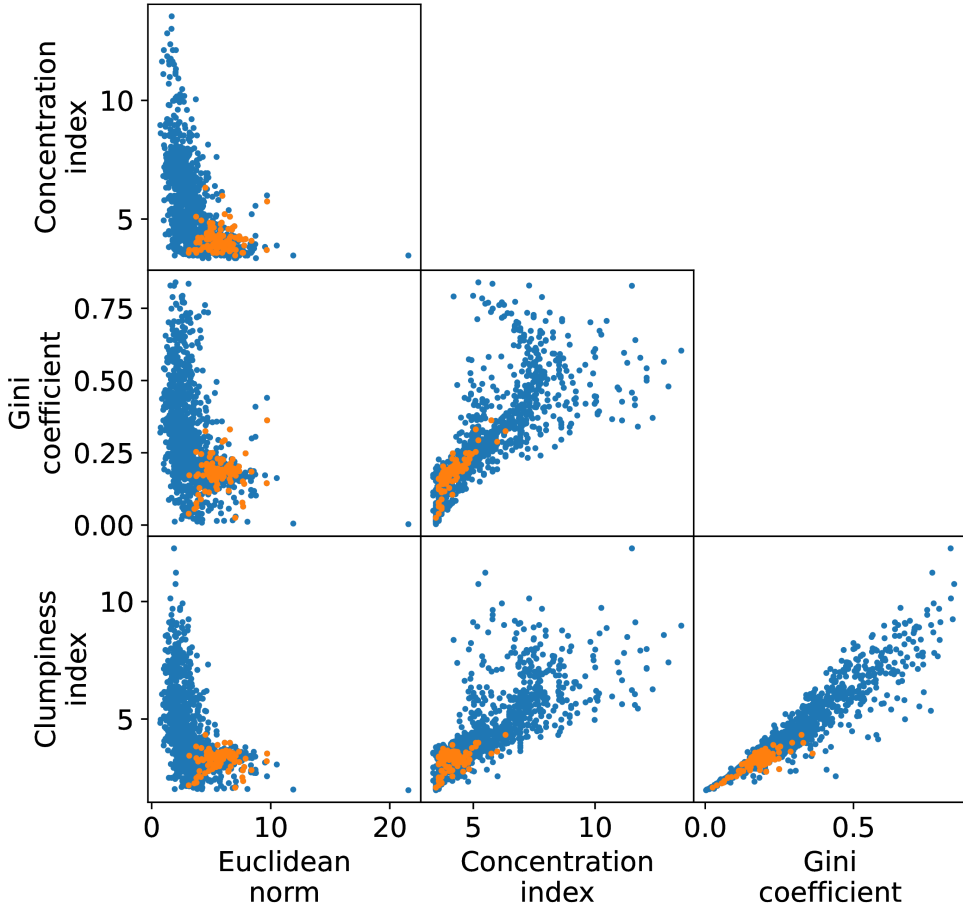


Figure 4.4: Morphology metrics for the AGN remnant candidates in the training set (orange) and the other sources in the training set (blue).

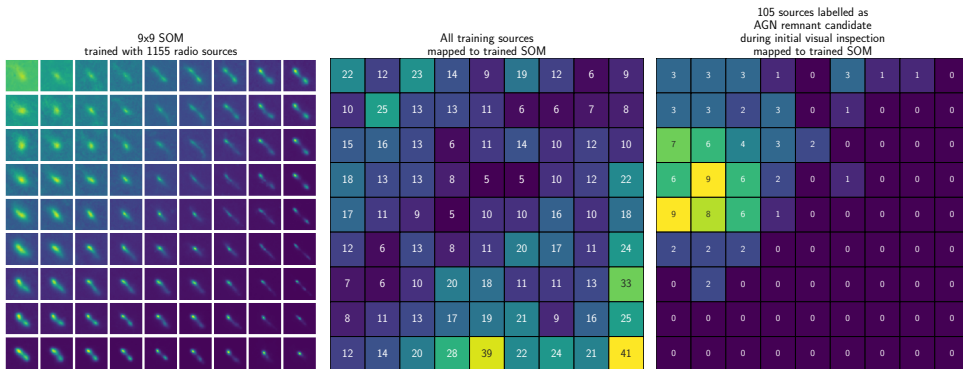


Figure 4.5: Results of SOM training. The first panel shows a 9×9 SOM trained with our training set. The second panel shows how all sources in the training set map to the SOM, and the third panel shows the subset of 'AGN remnant candidates' from the training set mapped to the SOM.



Figure 4.6: Results of compressed SOM training. Same as Fig. 4.5 except that the trained SOM is compressed to a 5×5 lattice. The second and third panels, respectively, show all training sources and the subset of ‘AGN remnant candidates’ mapped to this compressed SOM.

The middle and right panels of Fig. 4.5 show the mapping of all training sources to the SOM and the mapping of the AGN remnant candidates to the SOM, respectively. The middle panel of Fig. 4.5 shows that the sources are distributed over the entire SOM, with the highest concentrations in the bottom-right area of the SOM lattice. The right panel of Fig. 4.5 shows that the AGN remnant candidates are mostly located in the top-left region of the SOM, even though this pattern is somewhat scattered due to the low bin counts throughout. This low-bin-count effect will be worsened during training, cross-validation, and testing of our classifier as that will entail mapping smaller samples of sources to the SOM.

The left panel of Fig. 4.6 shows our compressed SOM, for which the same observations hold as those for the SOM in Fig. 4.5. However, additionally, we see that the mapping of the radio sources within this compressed SOM is more robust: the smaller number of neurons causes the distribution over the SOM of the low number of AGN remnant candidates to be less affected by small bin counts, especially during cross-validation. The AGN remnant candidate mappings to the compressed SOM in the middle panel of Fig. 4.6 show a clearer pattern: the AGN remnant candidates are mostly located in the diagonal band going from the top left to the bottom right with most mapped to the centre of the SOM, while the non-candidates are uniformly located, with most mapped to the bottom-right area of the SOM.

In our earlier work (Mostert et al., 2021), multiple neurons were needed to represent sources with similar morphology but with different apparent sizes. Figure 4.6 shows that the source-size rescaling that we introduced at the data pre-processing stage (Sect. 4.3.1) of this work reduces the size degeneracy in the resulting SOMs.

We note that there is no intrinsic significance to the absolute location of a neuron on the lattice. If one trains multiple SOMs with the same training data, similar neurons will always end up relatively close together on the lattice, but the absolute location of these neurons may change.

Quantitatively, the SOM training was monitored by logging the average quantisation error (AQE) and the topological error (TE). The AQE (Kohonen, 2001) is defined as taking the average over the summation of all the values of the Euclidean norm from each image in the dataset to its corresponding best-matching neuron. A lower AQE indicates that the SOM is a better representation of the data. The TE (Villmann et al., 1994) is defined as the percentage of image cutouts in our data

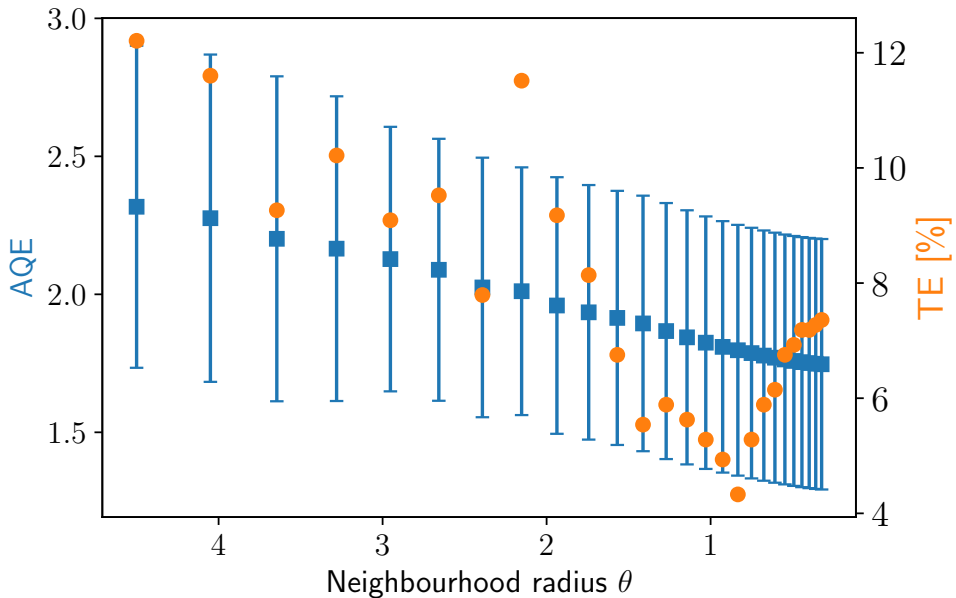


Figure 4.7: Quantitative training metrics for the SOM in Fig. 4.5, showing the steadily decreasing AQE on the left y-axis and the TE on the right axis.

4

for which the second best-matching neuron is not a direct neighbour of the best-matching neuron, where ‘direct neighbour’ is defined as all eight neighbours for neurons on a rectangular SOM lattice. TE is a measure for the coherence of the SOM, where lower TE signifies better coherency. With each SOM training update, one tries to lower AQE while keeping TE at acceptable values (arbitrary thresholds below 10% are common). Similar to the trends of the metrics in Mostert et al. (2021), Fig. 4.7 shows a steadily declining AQE and a TE that declines initially, reaches a minimum, and then climbs up again. In Mostert et al. (2021), training was stopped when the per-epoch decline of AQE dropped below 1%, while in the present work, with the particular training set that we use, we noticed that there is room to continue training up to 25 epochs to achieve a slightly lower AQE with a TE that is still solidly below 10%.

4.4.3 Resulting hyperparameters for the trained random forest classifier

In the cross-validation phase of training our RF, the grid search reveals that a maximum tree depth of 4, a maximum feature-ratio of 0.4, and a class weight of 0.16 for our majority class (and 0.84 for our minority class) are good hyperparameter choices for our RF when optimising for F_2 -score.

As the size of our training set is modest, we tested whether or not the predictions of our classifier are significantly better than chance. We did so by randomly permuting the labels in a five-fold stratified cross-validation process, and assessed how well our classifier is able to fit the features it receives to predict the noisy (permuted) labels. This allowed us to calculate the empirical p-value against the null hypothesis that our features and labels are independent. As the p-value represents the fraction of the permuted datasets where our trained model performed as well as or better than in the original

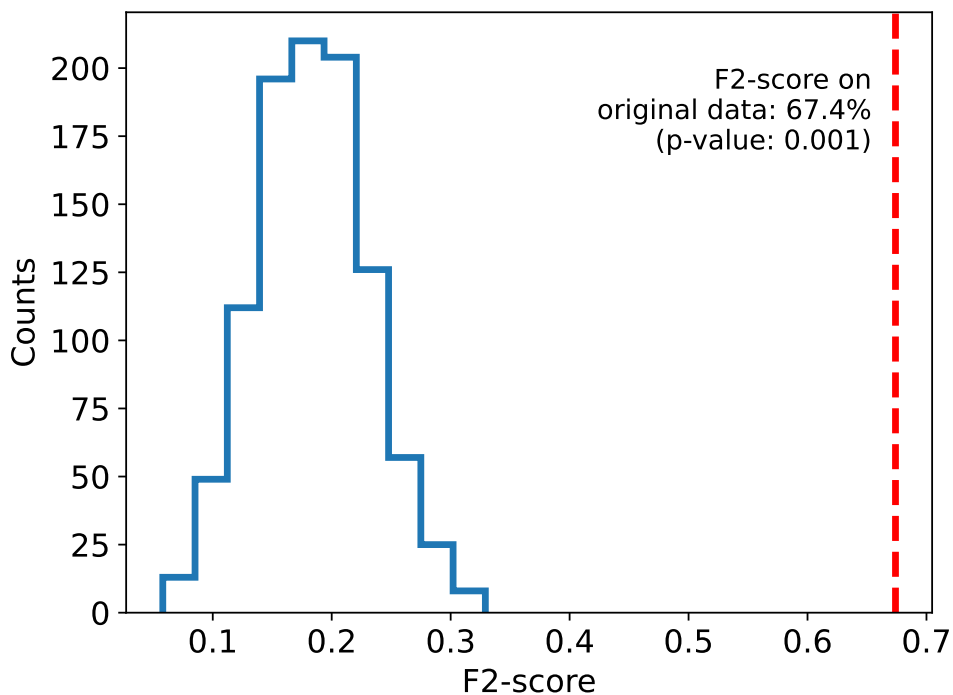


Figure 4.8: Permutation test scores. The red dashed line indicates the cross-validated F_2 -score of our RF classifier on the training data. The blue histogram shows the cross-validated F_2 -score of our RF classifier on the training data when the corresponding labels are randomly permuted. We repeated the shuffling of the labels and cross-validated score assessment 1000 times with different random seeds. The difference between the scores for the permuted data and those for the original data indicates the significance of our trained model.

Table 4.2: Normalised impurity-based importance of each feature in the training set according to RF. Higher values indicate higher importance.

Feature	Importance
Remnants ratio per SOM neuron	0.19
Remnants per SOM neuron	0.18
Total/peak flux	0.17
Euclidean norm	0.16
Clustered Haralick ratio	0.1
Clumpiness index	0.08
Concentration index	0.05
Major/minor axis	0.04
Gini coefficient	0.04

data, a high p-value (> 0.05) indicates a lack of dependency between the features and our labels or a badly trained model, and conversely a low p-value indicates the existence of a dependency and a well trained-model. Figure 4.8 shows the results of this permutation test. With a p-value of 0.001 for the F_2 -score of our model, we conclude that our features and labels are not independent and that our classifier is able to model a significant part of the mapping from our features to our labels.

4.4.4 Feature importance

We took two approaches to assess which (morphological) features contribute most to the predictive power of our classifier. The first is ‘mean decrease impurity feature importance’ (Breiman, 2001; Louppe, 2014), also known as the Gini importance (not to be confused with the Gini coefficient). Features used at the base of a decision tree contribute to the predicted label of a larger number of radio sources than features near the leaves of a decision tree. The relative depth of a feature in a tree of the RF can therefore be used as a proxy for its importance in predicting our class labels. We have to be cautious in interpreting these values, as impurity-based feature importance can be misleading for features with a high cardinality, favouring (continuous) features that can take on a high number of unique values over (discrete/ordinal) features with a low number of unique values (Strobl et al., 2007). In our case, the only ordinal feature is the ‘remnants per SOM neuron’. Furthermore, the impurity-based importance is based on the training set and features that the RF uses to overfit will also show up as important.

In our second approach, we test the importance of the different features that we use by means of feature permutation. Feature permutation importance is calculated by comparing the accuracy of our classifier with the original train or test data to the accuracy attained when the values of one of the features in the data is randomly shuffled. This process is repeated for all features and the results are normalised. We have to keep in mind that, using feature permutation, correlated features are shown as less important, because part of the performance loss of one permuted correlated feature is negated by the corresponding unpermuted correlated feature. In our case, the SOM-features are correlated, and from Fig. 4.4.1, the Gini coefficient, clumpiness index, and concentration index are also correlated.

Table 4.2 shows the results of the impurity-based feature importance and Fig. 4.9 the results of the feature permutation importance. It is reassuring to see that the two different feature importance methods point towards the same conclusion. Both approaches indicate that the importance of our

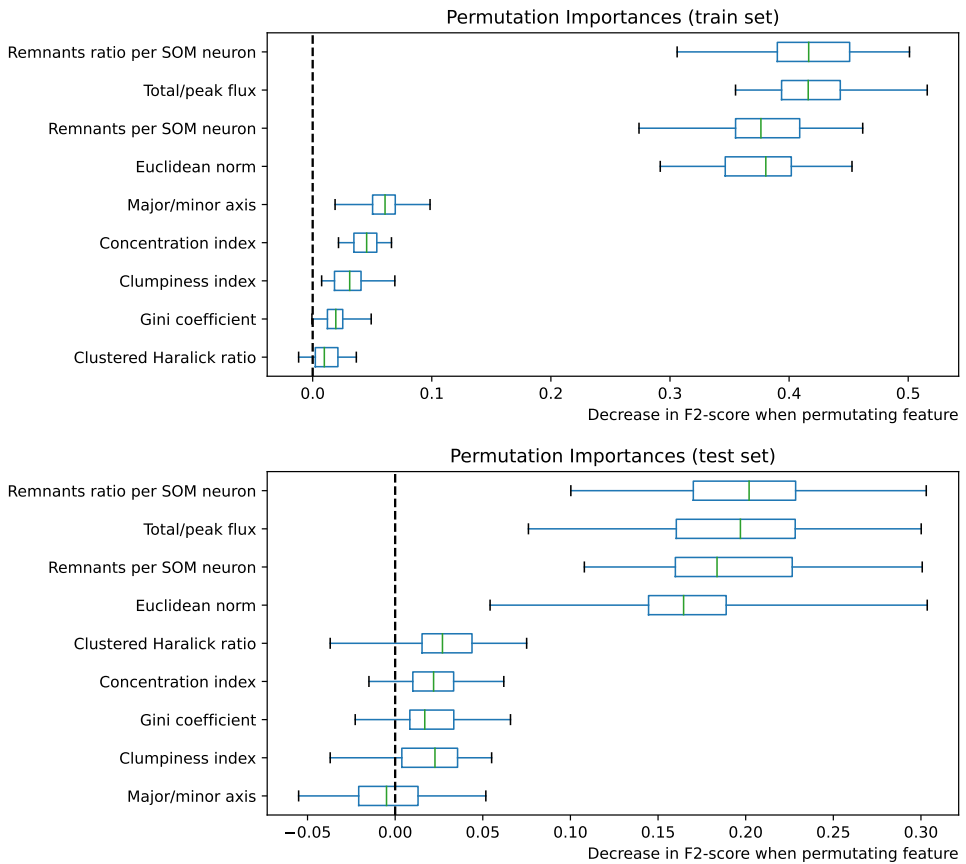


Figure 4.9: Importance of each feature permutation for our training set (top panel) and our test set (bottom panel) with respect to the F_2 -score. A higher F_2 -score indicates better predictive performance. The more important a feature is to the RF, the more the F_2 -score decreases after permuting this feature.

features can be roughly divided into two groups: the ‘major-to-minor axis ratio’, the ‘Haralick-derived features’, and the ‘features borrowed from optical galaxy morphology’ account for only a few per cent of the predictive power of the model, while the ‘SOM-derived features’ and the ‘total-to-peak flux ratio’ have the greatest influence. Taking a closer look at Fig. 4.9, we see the feature permutation importance for both the training and the test set. The test set values indicate which features are actually important in predicting the right class label. High positive values for features in the training set that are less positive or even negative for the test set indicate that the classifier uses these features to overfit on the training data. Looking at the difference between the training and test set values, most features are used by the RF to improve the model but also marginally overfit on the training set. Removing the ‘major-to-minor axis ratio’ feature might even marginally improve the resulting F_2 -score on the test set.

4.4.5 Random forest classifier performance

As we look for a small number of AGN remnant candidates in a large sample of yet-to-be inspected sources, isolating all candidates in the positive predictions is our most important objective for the classifier. In classification terms, this means that a ‘false negative’ —predicting that a radio source is not a candidate while it is— is less desirable than a ‘false positive’ —predicting that a radio source is a candidate while it is not. We can change the prediction threshold of our trained classifier (which is set to 0.5 by default) to generate fewer false negatives at the cost of more false positives. We change the prediction threshold to 0.25, as this is the point where the recall for our candidates is one hundred per cent. Although we focus on maximising the F_2 -score during training, resulting in a final F_2 -score of 0.71 on the train set and 0.46 on the test set, we also provide a summary of more common performance metrics in Table 4.3, thereby enabling easy comparison to other classifiers. The performance of the resulting classifier on our hold-out test can also be deduced from the corresponding confusion matrix shown in Fig. 4.10. The confusion matrix shows that for the 1,173 radio sources from our test set, we correctly discard 860 sources (true negatives, see Fig. 4.11), we correctly label 45 sources as candidates (true positives, see Fig. 4.12), we discard 0 candidates (false negatives), and we label 268 sources that have not yet been inspected as candidates (false positives, see Fig. 4.13). Thus, if our hypothesis is true and we verify that upon visual inspection the (yet-to-be-inspected) false negatives do indeed turn out to contain far more AGN remnant candidates than the (yet-to-be-inspected) true negatives, we can reduce the number of yet-to-be-inspected sources from 1,173 to $268 + 45 = 313$, a reduction of 73%. By accepting a non-zero number of false negatives, this reduction can be increased.

To put these numbers in perspective, we can estimate the absolute number of radio sources that would require visual inspection for the yet-to-be-completed LoTSS. Assuming that the completed LoTSS will cover 85% of the northern hemisphere (as suggested by Shimwell et al. (2022)), and extrapolating the source numbers from the preliminary LoTSS-DR2 catalogue,⁶⁶ we expect to find roughly 45 000 sources⁶⁷ > 60 arcsec in the complete LoTSS, for which our method would reduce visual inspection to roughly 12 000 sources.

⁶⁶The preliminary LoTSS-DR2 catalogue, which covers 27% of the northern sky, contains 14,287 sources > 60 arcsec.

⁶⁷or fewer, as the noise levels for further data releases are expected to be higher due to observing closer to the galactic plane.

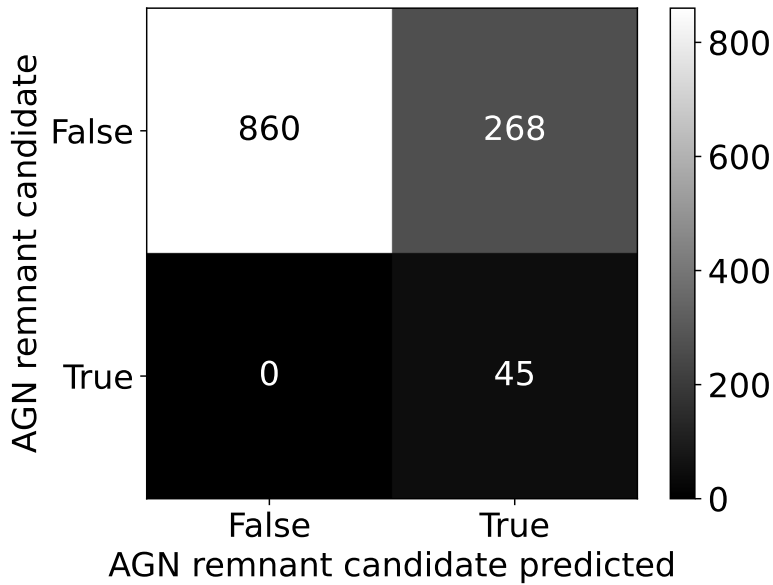


Figure 4.10: Confusion matrix for our trained classifier applied to the test set, using a prediction threshold of 0.25.

Table 4.3: Performance of trained RF on the test set with prediction threshold set to guarantee full recall of the AGN remnant candidates.

	precision	recall	f1-score	support
non-candidate	1.00	0.76	0.87	1128
remnant candidate	0.14	1.00	0.25	45
weighted average	0.97	0.77	0.84	1173
accuracy	0.77			1173

True negatives (less likely candidates)

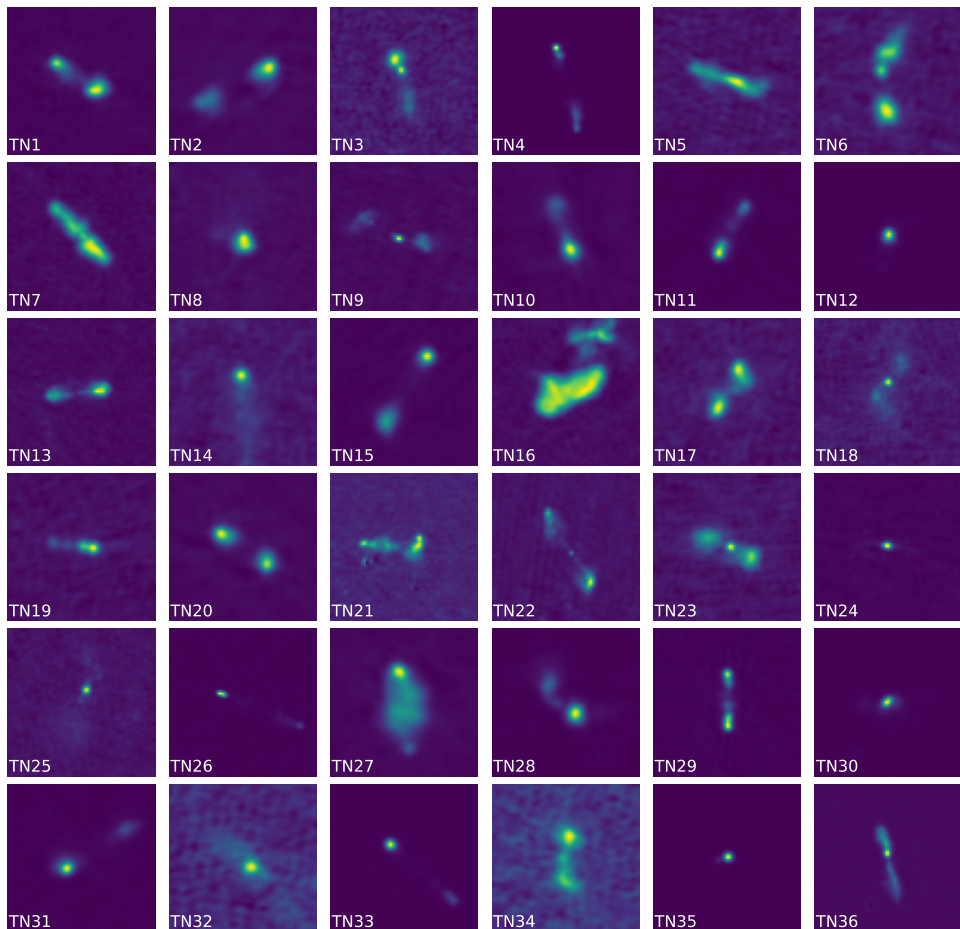


Figure 4.11: Randomly picked examples of true negatives from the test set. True negatives are yet-to-be inspected radio sources for which the model-predicted label is 'non-candidate'. As our trained model is able to distinguish these sources from the sources that were labelled 'AGN remnant candidate' during initial visual inspection, these sources are less likely to be labelled 'AGN remnant candidate' upon subsequent visual inspection.

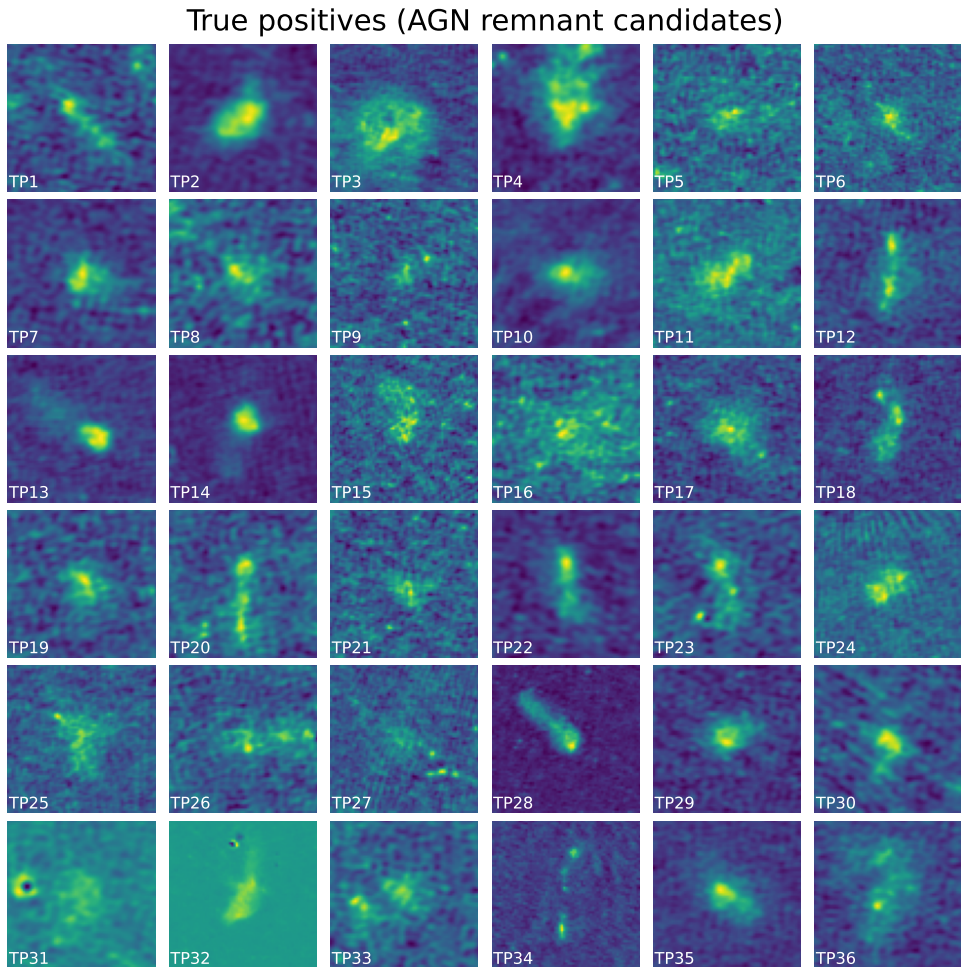


Figure 4.12: Randomly sampled examples of true positives from the test set. True positives are radio sources for which our model-predicted label matches the ‘AGN remnant candidate’ label from the initial visual inspection.

False positives (more likely candidates)

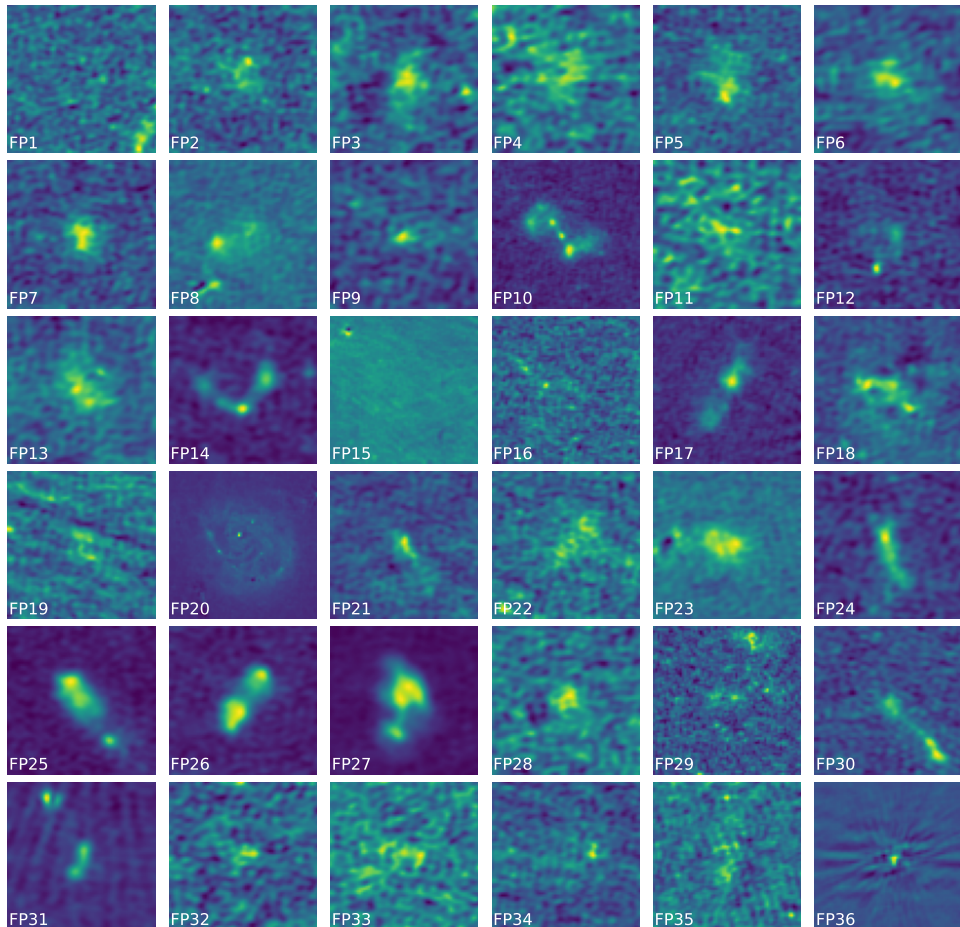


Figure 4.13: Randomly sampled examples of false positives from the test set. False positives are not yet inspected radio sources for which the model-predicted label is 'AGN remnant candidate'. False positives indicate that our trained model cannot distinguish these sources from the sources labelled 'AGN remnant candidate' during the initial visual inspection. Therefore, these false positives are *more likely* to be labelled 'AGN remnant candidate' upon subsequent visual inspection.

4.5 Discussion

In this section, we discuss the influence of the different features on classification and estimate the percentage of ‘AGN remnant candidates’ in the positive and negative predictions of our classifier and discuss the implications thereof. We also discuss the use and limitations of our methods, the added value of multi-frequency information, and future prospects.

4.5.1 Feature saliency

Proctor (2016), who used a decision tree classifier in combination with six source-catalogue-derived features to find giant radio galaxy candidates, suggested that including more features might allow better isolation of the sought-after radio sources. We observe that for AGN remnant candidates, SOM-derived morphological features can indeed complement source-catalogue-derived features. Table 4.2 and Fig. 4.9 show that the SOM-derived features and the total-to-peak flux ratio are the most important features. Specifically, the lower panel of Fig. 4.9 shows that the other features are roughly ten times less important to our RF classifier.

We suggest two explanations for the weak influence of the concentration index, the clumpiness index, the Gini coefficient, and the Haralick clusters. First, we can observe from Fig. 4.4 that these features are correlated, which reduces their importance. Second, an inspection of examples of sources and their feature value (see Appendices 4.A and 4.B) show that the noise around the central source and image artefacts from neighbouring sources strongly influence (or even dominate) the value of these extracted features. For high signal-to-noise-ratio sources, we could have negated much of the noise using sigma-clipping during pre-processing. Second, as we are concerned with sources that by definition have a relatively low surface brightness, we experimented with different (low) levels of sigma-clipping but we did not find a satisfactory compromise between loss of (diffuse) signal and noise removal, and therefore opted not to use sigma-clipping. The strong influence of the SOM features, despite the presence of the noise and emission from neighbouring sources, might be explained by the fact that the neurons in the SOM (each of which can be regarded as a tiny cluster) are suspended in a two-dimensional lattice, allowing greater separation of both morphology and relative noise levels.

4.5.2 Estimating the number of ‘AGN remnant candidates’ after model prediction

Given that most sources in our training and test set have not yet been visually inspected, we seek to know whether or not, upon drawing a source from the subset of sources for which our model prediction is positive (it predicts an ‘AGN remnant candidate’ label), the visual inspection will show that the source is indeed an ‘AGN remnant candidate’. Furthermore, we also want to answer this question for sources for which the model prediction is negative (it predicts a ‘non-candidate’ label). These answers allow us to estimate how many ‘AGN remnant candidates’ we expect to find upon visually inspecting all the positive model predictions, and how many we miss by not visually inspecting the negative model predictions.

To answer these questions, we randomly sampled (without replacement) and visually inspected 36 sources from the 268 sources in our test set with a positive prediction that have not yet been visually inspected (the ‘false positives’ in our test set). Similar to the initial visual inspection, we

performed visual inspection by looking at the LoTSS-DR2 radio emission of the source, its surroundings, and the overlapping optical data from Pan-STARRS1. Upon visual inspection, whereby we aimed to confirm the diffuse amorphous emission and the absence of compact components in the radio and the absence of SF in the optical, we found that 7 of the 36 draws can be labelled as ‘AGN remnant candidate’. From this experiment, plus the knowledge that 45 sources in the set of positive predictions were already assigned the label ‘AGN remnant candidate’ during initial visual inspection (Sect. 4.2), we know that the random variable,

$$p^*_1 := \frac{45}{313} + \frac{7}{36} \cdot \frac{268}{313} = 31\%, \quad (4.6)$$

is an unbiased estimator of the probability of drawing a source among positive predictions that turns out to have the label ‘AGN remnant candidate’ after visual inspection. Furthermore, we estimate the variance on p^* to be

$$W := \frac{268^2}{(313^2 \cdot 36^2)} \cdot 36 \cdot \frac{7}{36} \cdot \left(1 - \frac{7}{36}\right) \cdot \frac{232}{268 - 1} = 0.003. \quad (4.7)$$

As the standard deviation is the square root of the variance, we retrieve the result: $p^*_1 = 31 \pm 5\%$. (See Appendix 4.C for the full derivation.)

Analogously, we randomly sampled (without replacement) and visually inspected 36 sources from the 860 sources in our test set with a negative prediction that have not yet been visually inspected (the ‘true negatives’ in our test set). After visual inspection, none of the draws were labelled as ‘AGN remnant candidate’. This leads to the unsatisfying estimator $p^*_2 = 0$, and so instead we can use the upper bound of a (conservative) confidence interval. We apply the rule of three and estimate with 95% confidence that fewer than 1 in $36/3 = 12$ sources in our negative predictions will be an ‘AGN remnant candidate’. These findings support our hypothesis that we can speed up the process of finding more AGN remnant candidates by visually inspecting only the positive predictions of our model that was trained to separate a relatively small sample of radio sources with given ‘AGN remnant candidate’ label from a larger sample of yet-to-be inspected radio sources.

As the fraction of positive predictions is $f_p = \frac{45+268}{45+268+860} = 0.27$, if we were to base a census on just the positive predictions of our classifier, we would likely conclude that $f_p \cdot 31 \pm 5\% = 10 \pm 2\%$ of all the sources with a projected linear size > 60 arcsec in LoTSS are AGN remnant candidates. This percentage is in agreement with the AGN remnant fractions $< 9\%$, $< 11\%$, and $4\%–10\%$, within samples of > 1 arcmin sources, reported respectively by Mahatma et al. (2018), Jurlin et al. (2020), and Quici et al. (2021).

4.5.3 Use and limitations

Our methodology can also be repeated for the classification of other morphological groups for which only a few hundred labelled samples are available, such as core-dominated FRIs, wide-angle tailed, or narrow-angle tailed (bent) AGN sources (Mingo et al., 2019). Likewise, in the case of a new calibration set of AGN remnant candidates (because of observational follow-up or new insights), the method explained in this work could simply be reapplied. Compared to a deep neural network, which might also be considered for finding similar sources but is more akin to a ‘black-box’ approach, our methods allow for more insight into which features are most important to the

classifier.

Below, we discuss two main limitations of our method. First, our classifier is trained using expert labels in a supervised fashion. As a result, human error in the labelling (known as ‘label noise’) will unavoidably negatively affect the results of our classifier. Label noise decreases classification performance and forces a model to be more complex than necessary to be able to fit the noisy labels (Frénay & Verleysen, 2013). Label noise negatively affects any type of classification system, but RFs have at least demonstrated (Breiman, 2001; Folleco et al., 2008) to be more robust for datasets with asymmetric label noise than the commonly used Naive Bayes (Murphy, 2012) or C4.5 (Quinlan, 1993) classifiers.

Second, the strength of our method depends on the quality of the radio source catalogue. For example, a catalogue entry centred on a single radio lobe that is erroneously not associated with its other radio lobe, will appear as a rather amorphous oblong shape and might erroneously be classified as an AGN remnant candidate.

4.5.4 Future prospects

To improve future results using the methodology presented in this manuscript, as discussed in Sect. 4.5.1, it is essential to mitigate the effect of the noise around the radio sources. A sensible improvement would be to switch the source-detection software from PyBDSF, which returns a source–model build-up by multiple Gaussians and works best for unresolved or slightly resolved sources, towards a flood-fill type of source-detection software like ProFound (Robotham et al., 2018), which returns a pixel-resolution mask for each source and works better for complex well-resolved sources (Hale et al., 2019). This would allow us to mask all pixels outside of the radio source’s mask before extracting our morphological features and thereby reduce the effect of the surrounding noise.

In future studies, authors could also attempt to improve our classifier by additional training with sources that were selected for visual inspection (labelling) using active learning. Active learning is a ‘human-in-the-loop’ practice in machine learning that is designed to improve a model by alternating between model prediction and human labelling, with the concept that by labelling samples for which the model is least certain, the model requires fewer labels to acquire the same performance as a model that does not use active learning (Settles, 2009). Walmsley et al. (2020) show the benefits of using active learning in the context of classifying optical images of galaxies. A personalised active learning framework like Astronomaly (Lochner & Bassett, 2021) could be used for this purpose.

Another path towards improving our classifier would be to build a ‘foundation model’ (e.g. Azizi et al., 2021; Bommasani et al., 2021; Walmsley et al., 2022b). A foundation model is a very large neural network (e.g. LeCun et al., 1989; Goodfellow et al., 2016) that achieves improved performance on its predictions by (self-supervised) pre-training on a very large corpus of related data.⁶⁸ The combined data from the Square Kilometre Array (SKA; Braun et al., 2015) pathfinder surveys⁶⁹ are perfect for creating a foundation model for these surveys and future large-scale sky surveys from the SKA. Such a model would improve the performance on multiple downstream tasks, such as

⁶⁸Self-supervised learning entails training a neural network using labels that are self-derived from the data (e.g. Chen et al., 2020; Zbontar et al., 2021), in the image domain that can, for example, be achieved by masking a part of the image and training the network to predict the values of these masked pixels.

⁶⁹LoTSS and its low-frequency equivalent LoLSS (de Gasperin et al., 2023), MeerKAT (Jonas & MeerKAT Team, 2016) international GHz tiered extragalactic exploration (MIGHTEE; Jarvis et al., 2016; Heywood et al., 2022), and the Evolutionary Map of the Universe survey (EMU; Norris et al., 2011) by the Australian square-kilometre-array pathfinder (ASKAP; Johnston et al., 2008).

source morphology classification and source-component association (e.g. [Mostert et al., 2022](#)).

4.6 Conclusions

Large samples of AGN remnants would lead to progress in the study of the life cycle of radio galaxies. The number of observed radio galaxies is high thanks to the current SKA-pathfinder surveys, but current remnant samples are still small, as AGN remnants are quite rare and identifying them based on their morphology through visual inspection is time-consuming. In this work, we present an automated way to determine which radio sources in the LOFAR Two-metre Sky Survey are likely AGN remnant candidates based on their radio morphology. We train and test an RF classifier using features extracted from 4,075 radio sources, 150 of which were determined to be AGN remnant candidates through visual inspection by Brienza et al. in prep.

First, we automated the extraction of a large variety of morphological features, from simple statistics such as width-to-height ratio, to the Gini coefficient, clumpiness, and concentration indices normally used in optical astronomy, soft-clustered Haralick features, and SOM-derived features. We introduced size-invariance and a compression technique that makes the SOM more robust to small sample sizes, a necessity given that we started from 150 AGN remnant candidates, which is a small number of labels upon which to build a classifier. Using a grid search and cross-validation, we evaluated the best hyperparameters to train our RF classifier with. As we prefer visually inspecting more radio sources than missing AGN remnant candidates, we chose our prediction threshold such that we sacrifice precision for a recall of 100% on the already labelled AGN remnant candidates.

We took steps towards the quantification of the, so far, subjective nature of the radio morphology of AGN remnant candidates. Through feature permutation, we find that with the chosen setup, SOM-derived features and total-to-peak flux ratio are the most influential features for predicting the right class label, while clustered Haralick-features, Gini coefficient, concentration index, clumpiness index, and major-to-minor axis ratio are less relevant.

Running our model on a new sky region of the ongoing LoTSS survey, we could visually inspect only the positive predictions, reducing the fraction of sources requiring visual inspection by 73%. Within this set of positive predictions, we expect $31 \pm 5\%$ to be ‘AGN remnant candidate’. Using this approach, we will miss the ‘AGN remnant candidates’ that still reside in the negative predictions, but we conservatively estimate that these constitute fewer than 8% of the negative predictions.

The results from the method presented in this paper improve if a larger sample of labelled sources is available, as the predictive power of the features grows with increasing sample size. We conclude that our method brings us closer to identifying complete samples of AGN remnant candidates in an automatic way. However, given our current precision and the large volume of radio sources in LoTSS-DR2 and upcoming SKA and SKA-pathfinder data releases, further steps are needed to reach the ultimate goal of fully automatic morphological selection of AGN remnant candidates.

4.A Examples of radio sources for varying concentration index, clumpiness index, and Gini coefficient values

To show what features of a radio image are picked up by three morphology metrics normally used to assess optical galaxy morphology, we show nine sources in two parts of the parameter space for the concentration index (Fig. 4.14), the clumpiness index (Fig. 4.15), and the Gini coefficient (Fig. 4.16). It is interesting to see that, given our dataset of > 60 arcsec radio sources, four of the nine sources with high concentration values that we randomly picked (Fig. 4.14) show a double-double radio galaxy morphology (DDRG; Schoenmakers et al., 2000), a sign of restarted AGN activity.

4.B Examples of sources from different Haralick clusters

To see what features of a radio image are picked up by the clustered Haralick features, Figs. 4.17, 4.18 and 4.19 show nine randomly sampled sources for each of the four clusters, plus the one ‘noise’-cluster.

4.C Derivation of statistical estimates

We trained our RF classifier to separate the objects with class label ‘AGN remnant candidate’ from the other objects for which we do not yet know their class label. Our test set (containing $N = 1173$ sources) can be divided into two subsets. Dataset 1 comprises all $N_1 = 313$ positive predictions (sources in the test set that the model predicts to be ‘AGN remnant candidate’), and dataset 2 comprises all $N_2 = 860$ negative predictions (sources in the test set that the model predicts to be ‘non-candidate’). Dataset 1 can be further split into three subsets: the $K_1^A = 45$ sources with known ‘AGN remnant candidate’ class label, the unknown number (U_1^A) of sources that would be labelled ‘AGN remnant candidate’ upon visual inspection, and the unknown number (U_1^N) of sources that would be labelled ‘non-candidate’ upon visual inspection. We know that $U_1^A + U_1^N = N_1 - K_1^A = 268$.

We want to know the probability that visual inspection reveals ‘AGN remnant candidate’ upon sampling a source from dataset 1:

$$p = (K_1^A + U_1^A)/N_1 = K_1^A/N_1 + U_1^A/N_1. \quad (4.8)$$

Because N_1 , K_1^A , and U_1^A are constants and not random variables, p is a constant too. To estimate this probability, we randomly sample (without replacement) and visually inspect $N_s = 36$ sources from dataset 1. The number of sampled sources that turn out to be ‘AGN remnant candidate’ after visual inspection will be called x . In this case $x \sim \text{Hypergeometric}(U_1^A + U_1^N, U_1^A, N_s)$, yielding the expected value:

$$\mathbf{E}[x] = N_s \cdot U_1^A / (U_1^A + U_1^N), \quad (4.9)$$

which can be rewritten as:

$$\mathbf{E}[x/N_s] = U_1^A / (U_1^A + U_1^N), \quad (4.10)$$

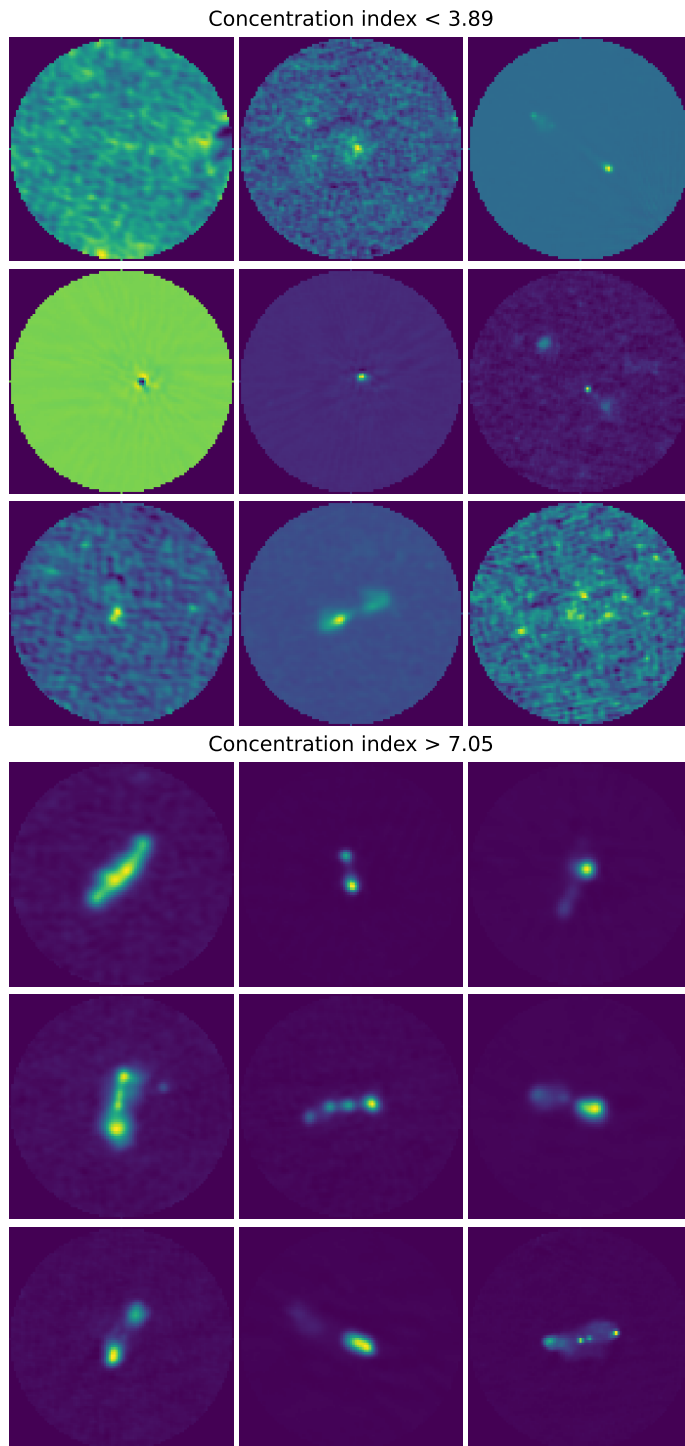


Figure 4.14: Randomly picked sources from the training set that showcase concentration index values below the 20th and above the 80th percentile.

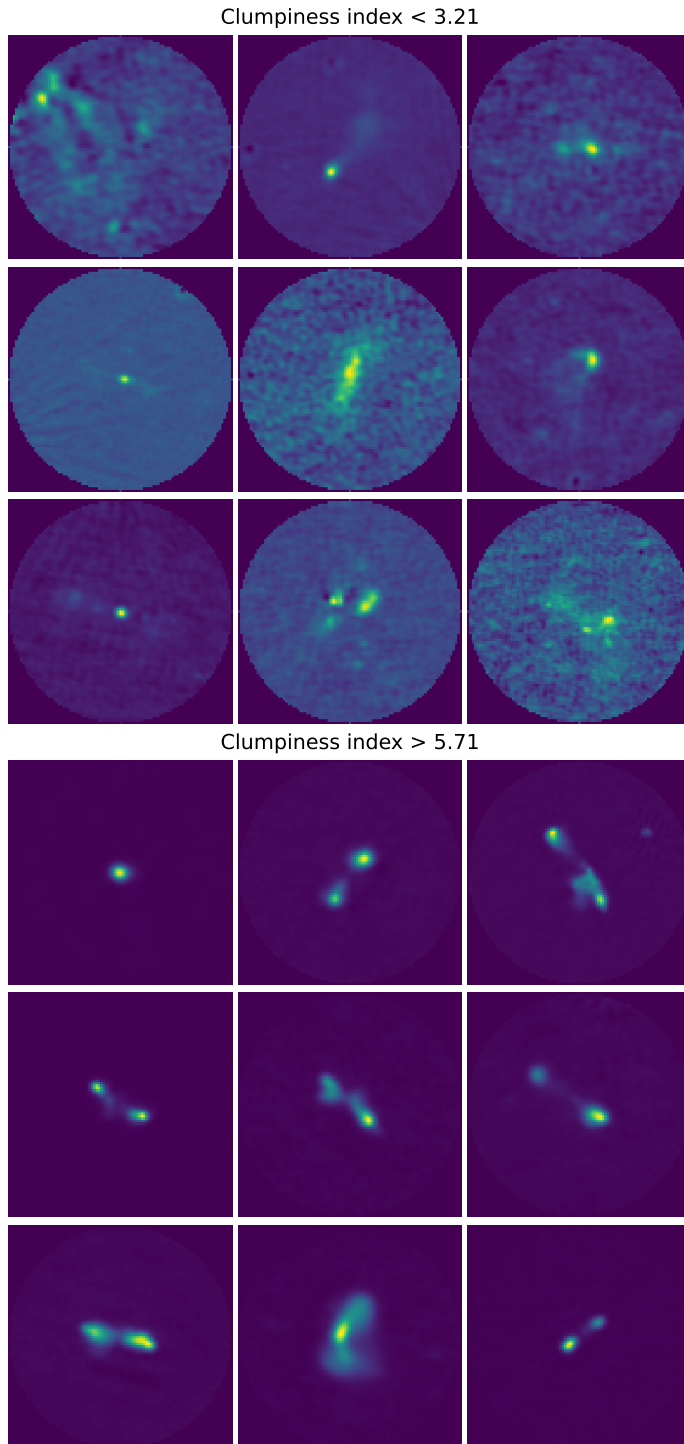


Figure 4.15: Randomly picked sources from the training set that showcase clumpiness index values below the 20th and above the 80th percentile.

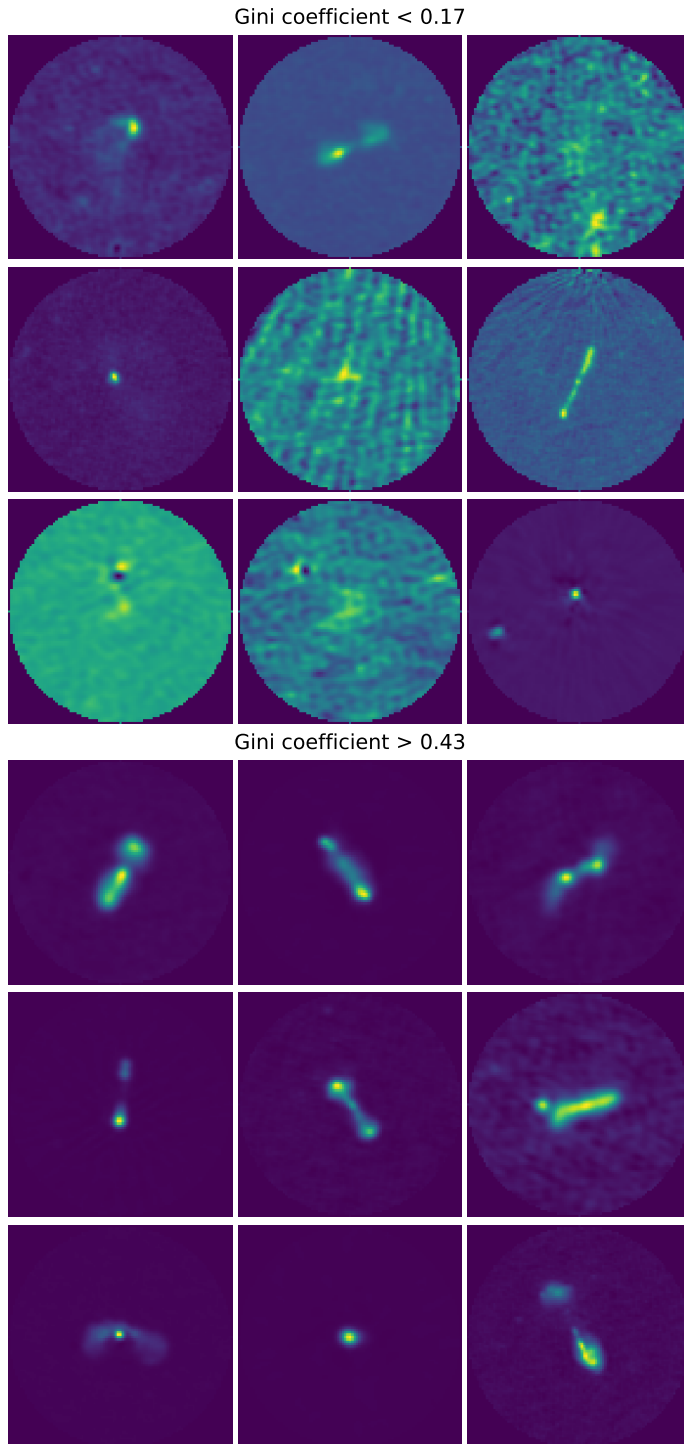


Figure 4.16: Randomly picked sources from the training set that showcase Gini coefficient values below the 20th and above the 80th percentile.

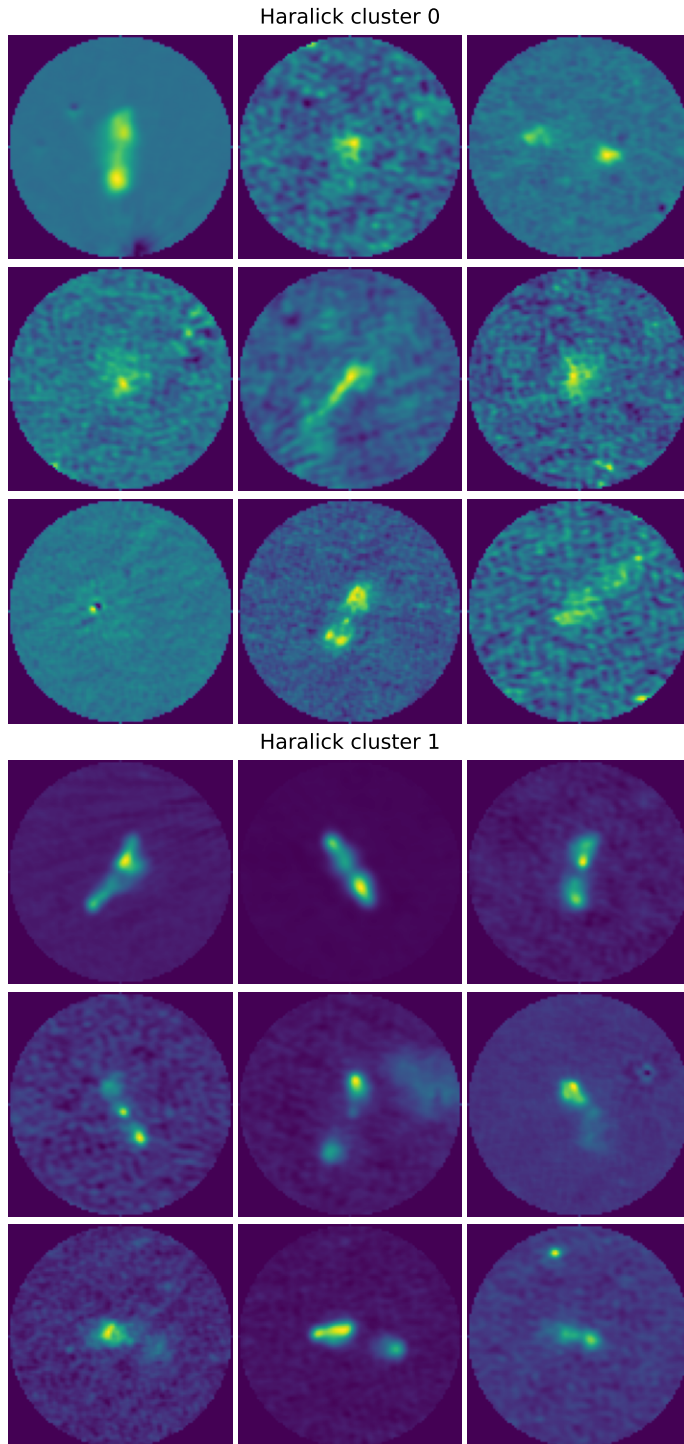


Figure 4.17: Nine randomly picked sources from the training set for each Haralick cluster.

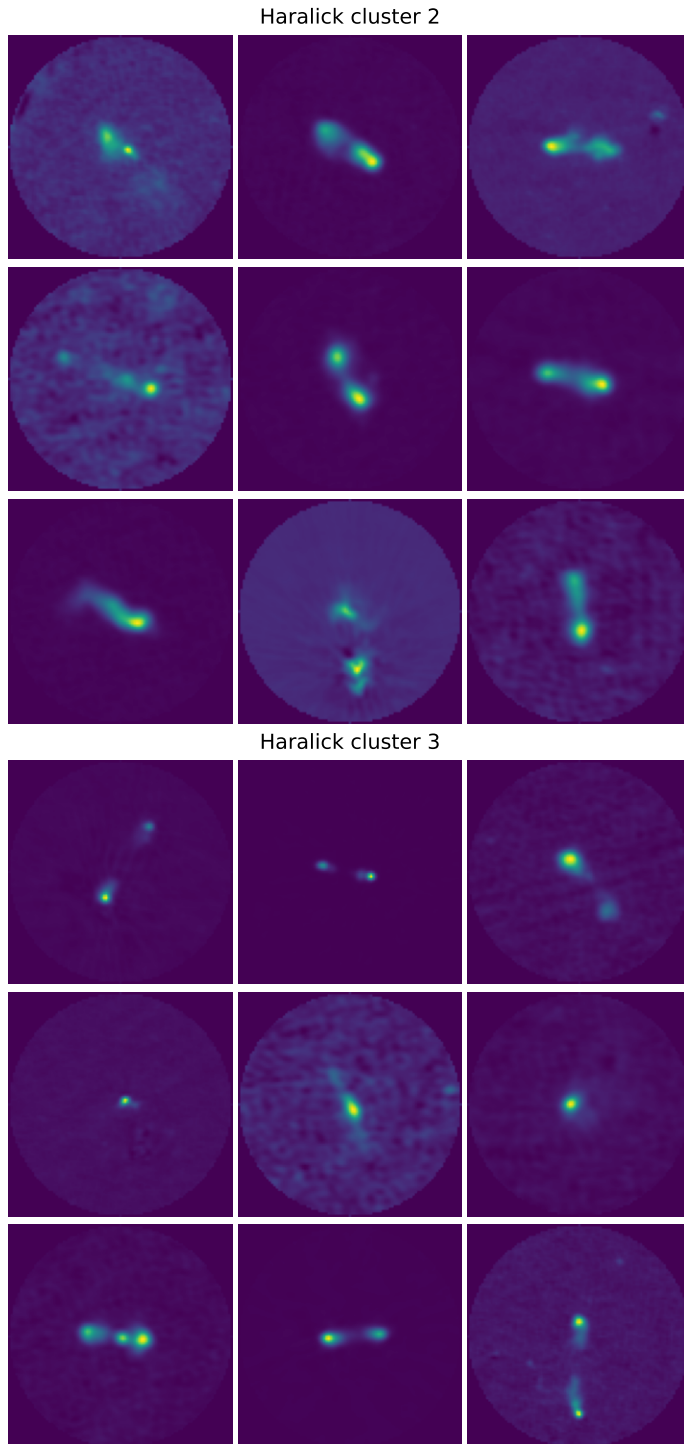


Figure 4.18: continued.

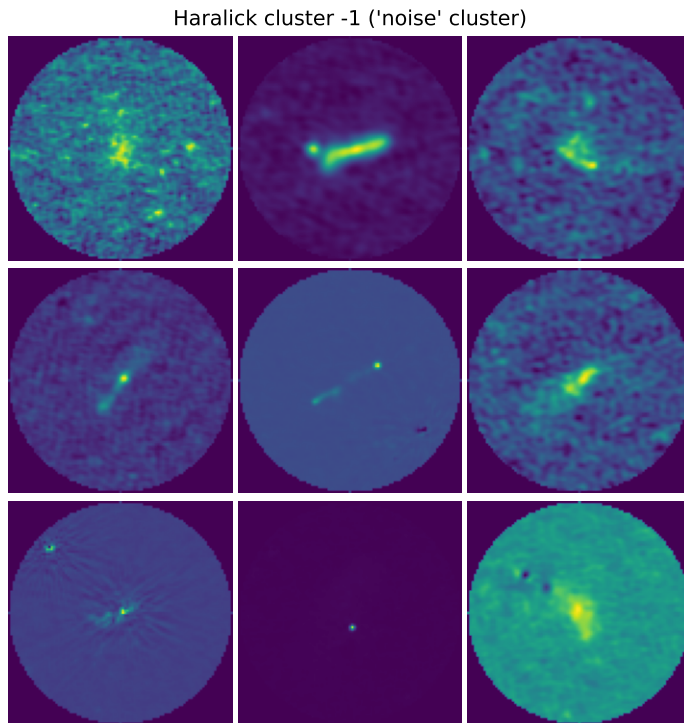


Figure 4.19: continued.

and as:

$$\mathbf{E}[x/N_s \cdot (U_1^A + U_1^N)/N_1] = U_1^A/N_1. \quad (4.11)$$

Combining Eqs. 4.8 and 4.11, we find that:

$$p = K_1^A/N_1 + \mathbf{E}[x/N_s \cdot (U_1^A + U_1^N)/N_1], \quad (4.12)$$

and it follows that the random variable

$$p^* := K_1^A/N_1 + x/N_s \cdot (U_1^A + U_1^N)/N_1 \quad (4.13)$$

is an unbiased estimator of p .

The variance of p^* is

$$\begin{aligned} \mathbf{V}[p^*] &= \mathbf{V}[K_1^A/N_1 + x/N_s \cdot (U_1^A + U_1^N)/N_1] \\ &= \mathbf{V}[x/N_s \cdot (U_1^A + U_1^N)/N_1], \end{aligned} \quad (4.14)$$

as the variance of constants is zero. Furthermore,

$$\mathbf{V}[p^*] = (U_1^A + U_1^N)^2 / (N_1^2 \cdot N_s^2) \mathbf{V}[x], \quad (4.15)$$

because $\mathbf{V}[aX] = a^2 \mathbf{V}[X]$ if a is a constant and X a random variable.

As we still consider the case $x \sim \text{Hypergeometric}(U_1^A + U_1^N, U_1^A, N_s)$, we know that

$$\begin{aligned} \mathbf{V}[x] &= N_s \cdot U_1^A / (U_1^A + U_1^N) \cdot U_1^N / (U_1^A + U_1^N) \\ &\quad \cdot (U_1^A + U_1^N - N_s) / (U_1^A + U_1^N - 1). \end{aligned} \quad (4.16)$$

We know from Eq. 4.10 that x/N_s is an unbiased estimator of $U_1^A / (U_1^A + U_1^N)$. Similarly $1 - x/N_s$ is an unbiased estimator of $U_1^N / (U_1^A + U_1^N)$. We can therefore estimate $\mathbf{V}[p^*]$ using random variable W :

$$W := (U_1^A + U_1^N)^2 / (N_1^2 \cdot N_s^2) V, \quad (4.17)$$

where

$$V := N_s \cdot x/N_s \cdot (1 - x/N_s) \cdot (U_1^A + U_1^N - N_s) / (U_1^A + U_1^N - 1). \quad (4.18)$$

We note that W is not an unbiased estimator of $\mathbf{V}[p^*]$.

Acknowledgements

The data and code used to produce the results and figures of this paper are available at: https://lofar-surveys.org/finding_agn_remnants.html. This research has made use of the Astropy ([The Astropy Collaboration et al., 2018](https://lofar-surveys.org/finding_agn_remnants.html)) and the scikit-learn ([Pedregosa et al., 2011a](https://lofar-surveys.org/finding_agn_remnants.html)) Python packages. The authors also acknowledge the usefulness of the PINK (<https://github.com/HITS-AIN/PINK>) and HaralickFeatures (<https://github.com/KushathaNtwaetsile/HaralickFeatures>) implementations. LOFAR data products were provided by the LOFAR Surveys Key Science project (LSKSP; <https://lofar-surveys.org/>) and were derived from observations with the

International LOFAR Telescope (ILT). LOFAR (van Haarlem et al. 2013) is the Low Frequency Array designed and constructed by ASTRON. It has observing, data processing, and data storage facilities in several countries, which are owned by various parties (each with their own funding sources), and which are collectively operated by the ILT foundation under a joint scientific policy. The efforts of the LSKSP have benefited from funding from the European Research Council, NOVA, NWO, CNRS-INSU, the SURF Co-operative, the UK Science and Technology Funding Council and the Jülich Supercomputing Centre. MB acknowledges financial support from INAF under the SKA/CTA PRIN “FORECaST”, from the agreement ASI-INAF n. 2017-14-H.O and from the PRIN MIUR 2017PH3WAT “Blackout”. M.S.S.L. Oei acknowledges support from the VIDI research programme with project number 639.042.729, which is financed by The Netherlands Organisation for Scientific Research (NWO). MJH acknowledges support from the UK STFC [ST/V000624/1]. LA is grateful for support from UK STFC via CDT studentship grant ST/P006809/1.

