Radio Galaxy Zoo: Machine learning for radio source host galaxy cross-identification

M. J. Alger[1,2]*, J. K. Banfield[1,3], C. S. Ong[2,4], L. Rudnick[5], O. I. Wong[6,3], C. Wolf[1,3], H. Andernach[7], R. P. Norris[8,9]

[1] Research School of Astronomy and Astrophysics, The Australian National University, Canberra, ACT 2611, Australia
[2] Data61, CSIRO, Canberra, ACT 2601, Australia
[3] ARC Centre of Excellence for All-Sky Astrophysics (CAASTRO)
[4] Research School of Computer Science, The Australian National University, Canberra, ACT 2601, Australia
[5] Minnesota Institute for Astrophysics, University of Minnesota, 116 Church St. SE, Minneapolis, MN 55455
[6] International Centre for Radio Astronomy Research-M468, The University of Western Australia, 35 Stirling Hwy, Crawley, WA 6009, Australia
[7] Departamento de Astronomía, DCNE, Universidad de Guanajuato, Apdo. Postal 144, CP 36000, Guanajuato, Gto., Mexico
[8] Western Sydney University, Locked Bag 1797, Penrith South, NSW 1797, Australia
[9] CSIRO Astronomy & Space Science, PO Box 76, Epping, NSW 1710, Australia

Accepted XXX. Received XXX

ABSTRACT

Radio source host galaxy cross-identification is the problem of determining the host galaxies of radio sources. While this is possible by hand, manual cross-identification is intractable for wide-area radio surveys like the Evolutionary Map of the Universe (EMU). We present a method for reducing the cross identification task to the standard machine learning task of binary classification. We apply our methods to the 1.4 GHz Australian Telescope Large Area Survey (ATLAS) observations of the *Chandra* Deep Field South (CDFS) and the ESO Large Area ISO Survey South 1 (ELAIS-S1) fields, cross-identifying them with the *Spitzer* Wide-area Infrared Extragalactic survey (SWIRE). We compare two sets of training data: expert cross-identifications of CDFS from the initial ATLAS data release and crowdsourced cross-identifications of CDFS from Radio Galaxy Zoo. We found a nearest neighbour approach outperforms our trained methods, likely resulting from ATLAS containing a low number of extended radio sources. Larger areas of sky containing more extended sources will be required for training methods such as ours to work with EMU. Our results show that the cross-identification accuracy of the predictor trained on Radio Galaxy Zoo cross-identifications is comparable to the predictor trained on expert cross-identifications, demonstrating the value of crowdsourced training data.

Key words: methods: statistical – techniques: miscellaneous – galaxies: active – radio continuum: galaxies – infrared: galaxies

1 INTRODUCTION

Next generation radio telescopes such as the Australian SKA Pathfinder (ASKAP; Johnston et al. 2007) and Apertif (Verheijen et al. 2008) will conduct increasingly wide, deep, and high-resolution radio surveys, producing large amounts of data. The Evolutionary Map of the Universe survey (EMU; Norris et al. 2011) using ASKAP is expected to detect over 70 million radio components, compared to the 2.5 million radio components currently known (Banfield et al. 2015).

An important part of processing this data is cross-identifying observed radio emission regions with observations of their host galaxy in surveys at other wavelengths. Cross-identification of the host with an extended radio souce can be a difficult task. Figure 1 illustrates the different radio emission regions that a host galaxy may have. Compact components are unresolved or point-like regions of radio emission separated from the host while compact radio sources have one compact component coincident with the host galaxy location. Extended radio sources consist of resolved single components or multiple (resolved or unresolved) components. The two most common classes of radio-loud sources, Fanaroff & Riley type I and type II (Fanaroff & Riley 1974), are examples of resolved radio sources. Small surveys containing a few thousand sources such as the Australia Telescope Large Area Survey (ATLAS; Norris et al. 2006; Middelberg et al. 2008) can be cross-identified manually, but this is impractical for larger surveys.

* Email: matthew.alger@anu.edu.au

One approach to cross-identification is crowdsourcing, where volunteers cross-identify radio components (both resolved and unresolved; see Figure 1) with their host galaxy. This is the premise of Radio Galaxy Zoo[1] (Banfield et al. 2015), a citizen science project hosted on the Zooniverse platform (Lintott et al. 2008). Volunteers are shown radio and infrared images and are asked to cross-identify radio sources with the corresponding infrared host galaxy. An explanation of the project can be found in Banfield et al. (2015). The first data release for Radio Galaxy Zoo will provide a large dataset of over 75 000 radio host cross-identifications and radio source morphologies (Wong et al., in prep). While this is a much larger number of visual cross-identifications than have been made by experts (e.g., Taylor et al. 2007; Gendre & Wall 2008; Grant et al. 2010; Norris et al. 2006; Middelberg et al. 2008) it is still far short of the millions of radio sources expected to be detected in upcoming radio surveys.

Automated algorithms have been developed for this cross-identification problem. Fan et al. (2015) developed a method of cross-identification using Bayesian hypothesis testing, fitting a three-component model to extended radio sources. This was achieved under the assumption that extended radio sources are composed of a core radio component and two lobe components. The core radio component is coincident with the host galaxy, so cross-identification amounts to finding the galaxy coincident with the core radio component in the most likely model fit. This method is easily extended to use other, more complex models, but it is purely geometric. The model does not incorporate other information such as the physical properties of the potential host galaxy. Additionally, there may be new classes of radio sources detected in future surveys like EMU which do not fit the model. Weston et al. (in press) developed a modification of the likelihood ratio method of cross-identification (Richter 1975) for application to ATLAS and EMU. This method does well on single (resolved or compact) radio sources with approximately 70 per cent success rate in the ATLAS fields, but does not currently handle extended multiple component radio sources (Norris 2017). The hope is that machine learning techniques can be developed for the cross-identification problem. Machine learning describes a class of methods that learn approximations to functions, and so if the cross-identification task can be cast as a machine learning problem, data sets such as that provided by Radio Galaxy Zoo can be generalised to work on data not seen by the original cross-identifiers.

We apply an approach from computer vision literature to the cross-identification task. This approach casts cross-identification as the standard machine learning problem of binary classification. We train our methods on expert cross-identifications and cross-identifications from Radio Galaxy Zoo. In section 2 we describe the data we use to train our methods. In section 3 we discuss how we cast the radio host galaxy cross-identification problem as a machine learning problem. In section 4 we present results of applying our method to ATLAS observations of the *Chandra* Deep Field South (CDFS) and in section 4.3 we apply the cross-identifiers trained on CDFS to the ESO Large Area ISO Survey South 1 (ELAIS-S1) field. Our data and code are available at https://radiogalaxyzoo.github.io/atlas-xid.

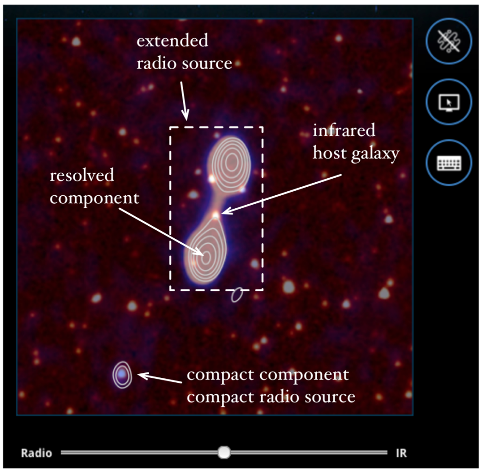


Figure 1. The Radio Galaxy Zoo tutorial image illustrating key definitions used throughout this paper. A colour version of this figure is available online.

Table 1. Catalogues of ATLAS/SWIRE cross-identifications for the CDFS and ELAIS-S1 fields. The method used to generate each catalogue is shown, along with the number of radio components cross-identified in each field.

Catalogue	Method	CDFS	ELAIS-S1
Norris et al. (2006)	Manual	784	0
Middelberg et al. (2008)	Manual	0	1366
Fan et al. (2015)	Bayesian models	784	0
Weston et al. (in press)	Likelihood ratio	3078	2113
Wong et al. (in prep)	Crowdsourcing	2460	0

2 DATA

We use data from the citizen science project Radio Galaxy Zoo (Banfield et al. 2015), the Australia Telescope Large Area Survey (ATLAS; Norris et al. 2006; Franzen et al. 2015), and the *Spitzer* Wide-area Infrared Extragalactic survey (SWIRE; Lonsdale et al. 2003; Surace et al. 2005). Radio Galaxy Zoo also includes data from the Faint Images of the Radio Sky at Twenty-Centimeters (FIRST; White et al. 1997). However, here we focus only on Radio Galaxy Zoo data from ATLAS and SWIRE.

2.1 ATLAS

ATLAS is a pilot survey for the EMU (Norris et al. 2011) survey, which will cover the entire sky south of +30 deg and is expected to detect approximately 70 million new radio sources. EMU will be conducted at the same depth and resolution as ATLAS, so methods developed for processing ATLAS data are expected to work for EMU. ATLAS is a wide-area radio survey of the CDFS and ELAIS-S1 fields at 1.4 GHz with a sensitivity of 14 and 17 µJy beam^{-1} on CDFS and ELAIS-S1 respectively. CDFS covers 3.6 deg^2 and contains 3034 radio components above a signal-to-noise ratio of 5. ELAIS-S1 covers 2.7 deg^2 and contains 2084 radio components above a signal-to-noise ratio of 5 (Franzen et al. 2015). The images of CDFS and ELAIS-S1 have angular resolutions of 16 by 7 and 12 by 8 arcsec respectively, with pixel sizes of 1.5 arcsec px^{-1}. Table 1 summarises catalogues that contain cross-identifications of radio components in ATLAS with host galaxies in SWIRE. In the present work, we train methods on CDFS and test these methods on both CDFS and ELAIS-S1. This ensures our methods are trans-

[1] https://radio.galaxyzoo.org

ferable to different areas of the sky observed by the same telescope as will be the case for EMU.

2.2 SWIRE

SWIRE (Lonsdale et al. 2003; Surace et al. 2005) is a wide-area infrared survey at the four IRAC wavelengths 3.6, 4.5, 5.8, and 8.0 µm (Lonsdale et al. 2003). It covers eight fields, including CDFS and ELAIS-S1. SWIRE is the source of infrared observations for cross-identification with ATLAS. SWIRE catalogues 221 535 infrared objects in CDFS and 186 059 infrared objects in ELAIS-S1 above a signal-to-noise ratio of 5.

2.3 Radio Galaxy Zoo

Radio Galaxy Zoo asks volunteers to cross-identify radio components with their infrared host galaxies. There are a total of 2460 radio components in Radio Galaxy Zoo sourced from ATLAS. These components are cross-identified by Radio Galaxy Zoo participants with host galaxies detected in SWIRE. A more detailed description can be found in Banfield et al. (2015) and a full description of how the Radio Galaxy Zoo Data Release 1 catalogue used in this work[2] is generated can be found in Wong et al. (in prep).

The ATLAS CDFS radio components that appear in Radio Galaxy Zoo are both unresolved and resolved radio components from the third data release of ATLAS by Franzen et al. (2015). Each radio component was fit with a two-dimensional Gaussian. Depending on the residual of the fit, more than one Gaussian may be fit to one region of radio emission. Each of these Gaussian fits is listed as a radio component in the ATLAS catalogue. The brightest radio component from the multiple Gaussian fit is called the 'primary component'. Each primary component found in the ATLAS DR3 component catalogue appears in Radio Galaxy Zoo. Non-primary components may appear within the image of a primary component, but do not have their own entry in Radio Galaxy Zoo. We will henceforth only discuss the primary components.

3 METHOD

3.1 Cross-identification as binary classification

We propose a two-step method for host galaxy cross-identification. Given a radio component, we want to find the corresponding host galaxy as a human would. The input is a $2' \times 2'$ radio image and a corresponding infrared image to match the size of the images used by Radio Galaxy Zoo. We make the assumption that each radio image represents a single, complex extended source. This limitation is discussed in section 3.2. The radio cross-identification task is then formalised as the computer vision problem of object localisation: given a radio image and an infrared image centred on a radio component, locate the host galaxy of the source containing the radio component.

A common approach to object localisation is to estimate the probability that each location in an image is coincident with the desired object, called a sliding window method (e.g. Rowley et al. 1996). Fixed-size windows of the image are taken centred on each pixel and we estimate the probability that the pixel is the location of the object. This task can be made more efficient if we assume that the host galaxy is always detected in the infrared. We then only consider windows centred on infrared sources, which we call 'candidate host galaxies'. This assumption usually holds in CDFS, except for a rare class of infrared-faint radio sources[3]. This defines a binary classification task where given a candidate host galaxy we compute the probability that it is a host galaxy. We refer to this binary classification task as the 'galaxy classification task'. To find

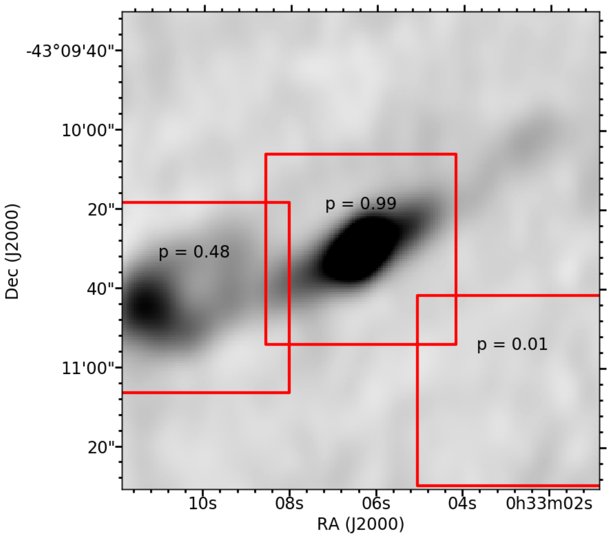


Figure 2. An example of localising the host galaxy of a radio source using our method. This image is a radio image from ATLAS and is centred on $\alpha = 00^{h}33^{m}06.36^{s}, \delta = -43^{\circ}10'30.1''$. Boxes represent 32 × 32 pixel windows centred on various locations in the image. The image contained in each window is used to represent the location that the window is centred on. The probabilities of each patch coinciding with the host galaxy would then be estimated by a classification model. The probabilities shown are for illustration only. In this example, the centre window of the image would be chosen as the location of the host galaxy, as the window centred on it has the highest probability.

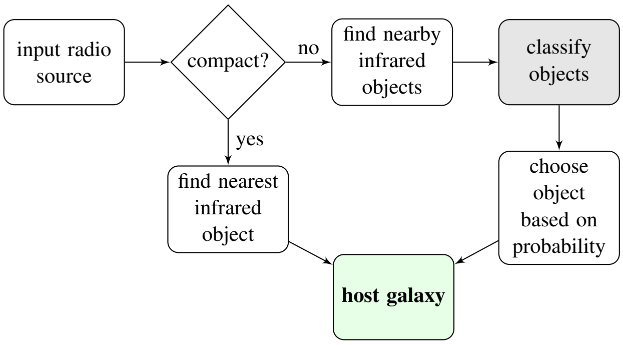

input radio source
compact?
no
yes
find nearby infrared objects
classify objects
find nearest infrared object
choose object based on probability
host galaxy

Figure 3. A cross-identification method employing a binary classifier. As input we accept a radio source. If the source is compact, we select the nearest infrared object as the host galaxy. If the source is resolved, we classify all infrared objects nearby within radius R and select the highest probability object as the host galaxy. The grey box is the classifier, which can be any binary classifier that outputs a probability.

[2] The Radio Galaxy Zoo Data Release 1 catalogue will only include cross-identifications for which over 65 per cent of volunteers agree. However, we use all volunteer cross-identifications for all sources.

[3] Norris et al. (2006) found 22 such radio sources in their sample of 784 bright ATLAS components in CDFS.

the host galaxy given a radio source, we can classify each galaxy within 1 arcmin of the source and select the host galaxy based on the probability output by a classifier. We refer to this as the 'cross-identification task'. This is a two-step task as we first solve the galaxy classification task without reference to any specific radio object, and then use these results to solve the cross-identification task. Our approach is illustrated in Figure 2.

Solving the galaxy classification task amounts to modelling a function f from the set of infrared sources $\mathcal{X}$ to the probability that an infrared source belongs to a binary class in $\mathcal{Y} = \{0, 1\}$:

$$f(x) \coloneqq \Pr(\mathcal{Y} = 1 \mid \mathcal{X} = x) \quad . \tag{1}$$

The space of infrared sources $\mathcal{X}$ needs to be encoded as a vector for the models we will use. We describe this in section 3.3. There are many options for modelling f. In this paper we apply three different models: logistic regression, random forests, and convolutional neural networks.

Note that $f(x)$ is independent of any particular radio object. The galaxy classification task aims to answer the general question of whether a given galaxy is the host galaxy of *any* radio component. The cross-identification task, however, attempts to cross-identify a *specific* radio component. It is unclear how to pass radio component-specific information into the classifiers. Instead, we assume that the probability that a given candidate host x is associated with a given radio component r is the probability that x is a host weighted by some radio-component specific value $g(x, r)$:

$$\Pr(x \text{ is host of } r) = f(x)g(x, r).$$

We then want to maximise $\Pr(x \text{ is host of } r)$ to solve the cross-identification task. One approach would be to select the candidate galaxy with the highest estimated class probability, i.e. take $g(x, r) = 1$ and let the host galaxy be the candidate galaxy x that maximises $f(x)$. This fails if there are multiple host galaxies within the $2' \times 2'$ image. An example of such an image is shown in Figure 4. Instead, we take g to be a Gaussian function of the angular separation of the candidate host from the radio component we want to cross-identify. The Gaussian essentially acts as a prior for the host galaxy location. The width of the Gaussian, σ, controls the influence of the Gaussian on the final cross-identification. When σ is small, our approach is equivalent to a nearest-neighbours approach where we select the nearest infrared object to the radio component as the host galaxy. In the limit where $\sigma \to \infty$, we maximise the probability output by the classifier as above. We take $\sigma = 30''$ as this was the best value found by a grid search.

We can improve upon this method by filtering out compact radio sources, which are much easier to cross-identify — the nearest SWIRE object may be identified as the host galaxy, or a more complex method such as likelihood ratios may be applied (see Weston et al. in press). A full cross-identification pipeline making use of this alongside a binary classifier is shown in Figure 3.

3.2 Limitations of our approach

We make a number of assumptions to relate the cross-identification task to the galaxy classification task:

- A radio image represents a whole, single radio source.
- The host galaxy of a radio component is within 1 arcmin of the component.
- The host galaxy of a radio component is closer on the sky to the radio component than the host galaxy of any other radio component.

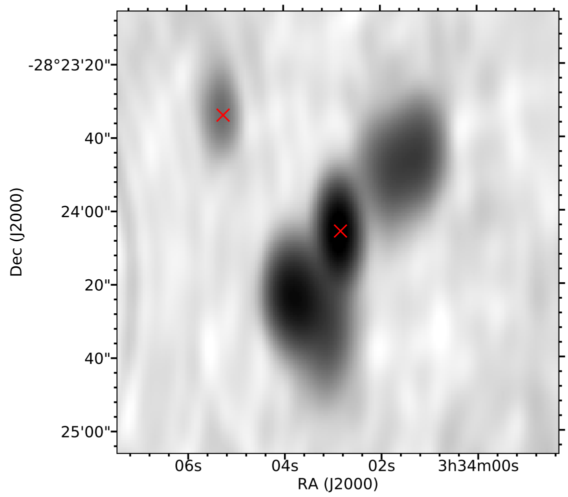


Figure 4. 2′-wide radio image centred on ATLAS3_J033402.87-282405.8C. This radio source breaks the assumption that there are no other radio sources within 1 arcmin of the source. Another radio source is visible to the upper-left. Host galaxies found by Radio Galaxy Zoo volunteers are shown by crosses.

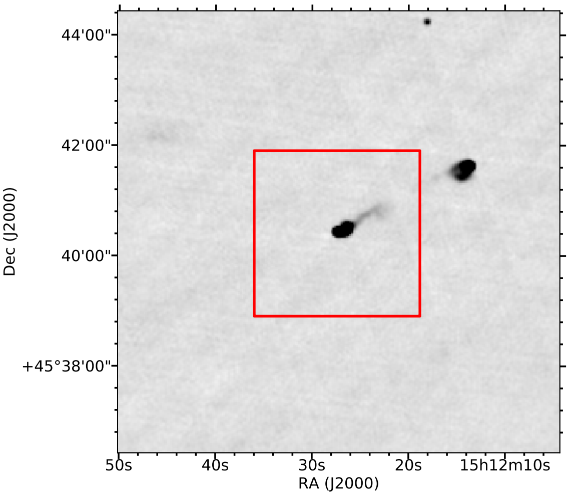


Figure 5. A 8′-wide radio image from FIRST, centred on FIRSTJ151227.2+454026. The 3′-wide red box indicates the boundaries of the image of this radio component shown to volunteers in Radio Galaxy Zoo. This radio source breaks our assumption that the whole radio source is visible in the chosen radius. As one of the lobes of the radio source is outside of the image, a volunteer (or automated algorithm) looking at the 3′-wide image may be unable to determine that this is a radio double or locate the host galaxy.

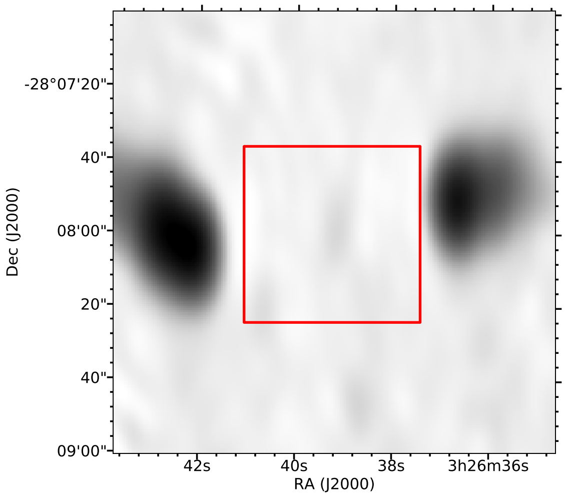


Figure 6. A radio image centred on $\alpha = 03^h26^m39.12^s$, $\delta = -28°07'58.80''$. This is an example of a radio source where the window centred on the host galaxy, shown as a rectangle, does not contain enough radio information to correctly identify the galaxy as the host.

- The host galaxy appears in the SWIRE catalogue.

These assumptions limit the effectiveness of our approach, regardless of how accurate our binary classification may be.

The main limitation is in choosing the image width and candidate search radius. If the radio image does not contain the whole source, then we are missing radio information useful for finding the host galaxy. If the radio image contains multiple sources, then we may be giving our classifiers irrelevant information. If the search radius is too small, we may not consider the host galaxy as a candidate. If the search radius is too large, we may consider multiple host galaxies (though this is somewhat mediated by the Gaussian multiplier). These kinds of size problems are difficult even for non-automated methods as radio sources can be extremely wide — for example, Radio Galaxy Zoo found a radio giant that spanned over three different images presented to volunteers and the full source was only cross-identified by the efforts of citizen scientists (Banfield et al. 2015). An example of a radio image where part of the radio source is outside the search radius is shown in Figure 5.

A related issue is that we need to choose a window size for the radio image representations of each infrared object. If this image is too small, radio emission may extend past the edges of the window, and it may be impossible to identify the galaxy as a host galaxy. If the image is too large, then too much information will be included and it will be difficult or computationally expensive to classify. We chose a window size of 32 × 32 pixels, corresponding to approximately $48'' \times 48''$ in ATLAS. This is shown as red rectangles in Figure 2 and Figure 6.

The assumption that the host galaxy of a radio component is closer on the sky to the radio component than any other host galaxy is required for the Gaussian distance weighting to correctly break ties.

We only need the assumption that the host galaxy appears in SWIRE to incorporate galaxy-specific features (section 3.3) and to improve efficiency. Our method is applicable even to infrared-faint radio sources by considering every pixel of the radio image as a candidate location as would be done in the original computer vision approach.

Our assumptions impose an upper bound on how well we can cross-identify radio sources. We estimate this upper bound in section 4.1.

3.3 Feature vector representation of infrared sources

Classification methods require that the inputs to be classified are represented by an array of real values called feature vectors. We thus need to choose a feature vector representation of our candidate host galaxies.

We use the SWIRE catalogue to find candidate hosts to classify. Candidates are represented by a a 32 × 32 pixel radio image from ATLAS, centered on the location of the candidate. The candidate host location and infrared colours encode almost all information we could gain from the infrared image, so we do not use the infrared image in this work.

We represent each candidate host as 1034 real-valued features. For a given candidate host, these features are:

- the 6 logarithms of the ratios of fluxes of the candidate host at the four IRAC wavelengths;
- the stellarity index of the host at both 3.6 and 4.5 µm;
- the flux of the host at 3.6 µm;
- the radial distance between the candidate host and the nearest radio component in the ATLAS catalogue; and
- a 32 × 32 pixel image from ATLAS (approximately $48''\times48''$), centred on the candidate host.

The infrared fluxes provide insight into the properties of the host galaxy of the radio emission. The 3.6 and 4.5 µm fluxes trace both galaxies with faint polycyclic aromatic hydrocarbon (PAH) emission and elliptical galaxies dominated by old stellar populations. The 5.8 µm flux selects galaxies where the infrared emission is dominated by non-equilibrium emission of dust grains (PAH destruction by the hard UV spectrum of AGN), while the 8.0 µm flux traces strong PAH emission at low redshift (Sajina et al. 2005). The stellarity index represents how likely the object is to be a star rather than a galaxy (Surace et al. 2005).

We use the pixels of each 32 × 32 radio image as independent features for all classifiers, with the convolutional neural network automatically extracting features that are relevant. Other features of the radio components may be used instead of the pixel values, but there has been limited research on extracting such features. Proctor (2006) describes hand-selected features for radio doubles in FIRST, and Aniyan & Thorat (2017) and Lukic et al. (in prep) make use of deep convolutional neural networks which automatically extract features as part of classification. A more comprehensive investigation of features is beyond the scope of this initial study, and is a good avenue for potential improvement in our pipeline.

3.4 Classifiers

We use three different classifiers as our binary classification model: logistic regression, convolutional neural networks, and random forests. These classifiers cover three different approaches to machine learning. Logistic regression is a probabilistic binary classification model. It is linear in the feature space and outputs the probability that the input has a positive label (Bishop 2006, Chap. 4). Convolutional neural networks (CNN) are a biologically-inspired

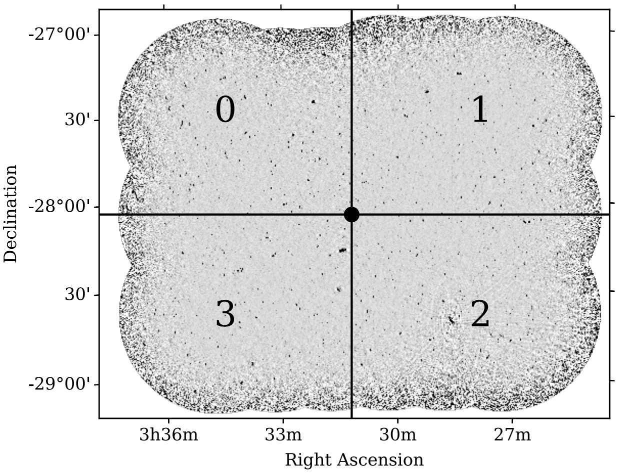


Figure 7. CDFS field training and testing quadrants labelled 0 – 3. The central dot is located at $\alpha = 03^{\mathrm{h}}31^{\mathrm{m}}12^{\mathrm{s}}, \delta = -28°06'00''$. The quadrants were chosen such that there are similar numbers of radio sources in each quadrant.

Table 2. Number of compact and resolved radio objects in each CDFS quadrant. Radio Galaxy Zoo (RGZ) has more cross-identifications than the expert catalogue provides as it uses a deeper data release of ATLAS, and so has more objects in each quadrant for training.

Quadrant	Compact	Resolved	Compact (RGZ)	Resolved (RGZ)
0	126	24	410	43
1	99	21	659	54
2	61	24	555	57
3	95	18	631	51
Total	381	87	2255	205

prediction model for prediction with image inputs. They have recently produced good results on large image-based datasets in astronomy (e.g. Dieleman et al. 2015, Lukic et al. in prep). Random forests are an ensemble of decision trees (Breiman 2001). It considers multiple subsamples of the training set, where each bootstrap subsample is sampled with replacement from the training set. To classify a new data point, the random forest takes the weighted average of all classifications produced by each decision tree.

Further details and background of these models are available in the supplement[4].

3.5 Labels

Converting the Radio Galaxy Zoo and Norris et al. (2006) cross-identification catalogues to binary labels for infrared objects is a non-trivial task. One problem is that there is no clear way to capture radio morphology information in binary classification of infrared objects. Another problem is that there is no way to indicate *which* radio object an infrared object is associated with, only that it is associated with *some* radio object. We ignore these problems for this paper and make the naïve assumption that any given search radius contains only one host galaxy. This allows us to assume that if a candidate host galaxy in the search radius of a component is not the host galaxy of that component, then it is not a host galaxy at all.

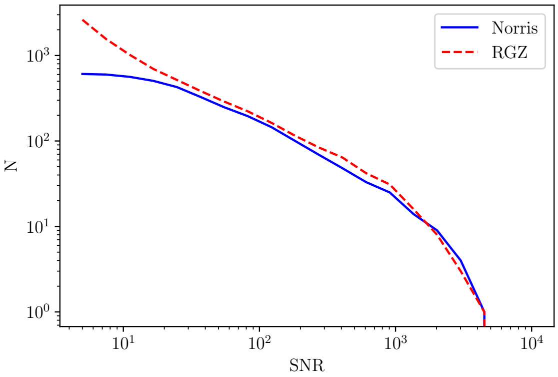


Figure 8. Number of radio components (N) in the expert (Norris) and Radio Galaxy Zoo (RGZ) training sets with different signal-to-noise (SNR) cutoffs.

We then generate positive labels from a cross-identification catalogue. We decide that if an infrared object is listed in the catalogue, then it is assigned a positive label as a host galaxy. We then assign every other galaxy a negative label. This has some problems — an example is that if the cross-identifier did not observe a radio object (e.g. it was below the signal-to-noise ratio) then the host galaxy of that radio object would receive a negative label. This occurs with Norris et al. (2006) cross-identifications, as these are associated with the first data release of ATLAS. The first data release went to a 5σ flux density level of $S_{1.4} \geq 200$ µJy beam^{-1} (Norris et al. 2006), compared to $S_{1.4} \geq 85$ µJy beam^{-1} for the third data release used by Radio Galaxy Zoo (Franzen et al. 2015). The labels from Norris et al. (2006) may therefore disagree with labels from Radio Galaxy Zoo even if they are both plausible. The difference in training set size at different flux cutoffs is shown in Figure 8. We train and test our classifiers on infrared objects within a 1 arcmin radius of an ATLAS radio object.

3.6 Experimental Setup

We trained binary classifiers on infrared objects in the CDFS field using two sets of labels. One label set was derived from Radio Galaxy Zoo cross-identifications and the other was derived from the Norris et al. (2006) cross-identification catalogue. We refer to these as the 'Radio Galaxy Zoo labels' and the 'expert labels' respectively. We divided the CDFS field into four quadrants for training and testing. The quadrants were divided with a common corner at $\alpha = 03^{\mathrm{h}}31^{\mathrm{m}}12^{\mathrm{s}}, \delta = -28°06'00''$ as shown in Figure 7. For each trial, one quadrant was used to draw test examples and the other three quadrants were used for training examples.

We further divided the radio components into compact and resolved. Compact components are cross-identified by fitting a 2D Gaussian (as in Norris et al. 2006) and we would expect any machine learning approach for host cross-identification to attain high accuracy on this set. A radio component was considered resolved if

$$\ln\left(\frac{S_{\mathrm{int}}}{S_{\mathrm{peak}}}\right) > 2\sqrt{\left(\frac{\sigma_{S_{\mathrm{int}}}}{S_{\mathrm{int}}}\right)^2 + \left(\frac{\sigma_{S_{\mathrm{peak}}}}{S_{\mathrm{peak}}}\right)^2}\ , \qquad (2)$$

[4] https://radiogalaxyzoo.github.io/atlas-xid/

where S_{int} is the integrated flux density and S_{peak} is the peak flux density (following Franzen et al. 2015).

Candidate hosts were selected from the SWIRE catalogue. For a given subset of radio components, all SWIRE objects within 1 arcmin of all radio components in the subset were added to the associated SWIRE subset. In the context of the galaxy classification task, we refer to SWIRE objects within 1 arcmin of a compact radio component as part of the 'compact set', and SWIRE objects within 1 arcmin of a resolved radio component as part of the 'resolved set'.

To reduce bias in the testing data due to the expert labels being generated from a shallower data release of ATLAS, a SWIRE object was only included in the test set if it was within 1 arcmin of a radio object with a cross-identification in both the Norris et al. (2006) catalogue and the Radio Galaxy Zoo catalogue.

Each classifier was trained on the training examples and used to predict labels for the test examples. The predicted labels were compared to the expert labels and the balanced accuracy was computed. Balanced accuracy is the average of the accuracy on the positive class and the accuracy on the negative class, and is not sensitive to class imbalance. The galaxy classification task has highly imbalanced classes — in our total set of SWIRE objects within 1 arcmin of an ATLAS object, only 4 per cent have positive labels. Only examples within 1 arcmin of ATLAS objects in the first ATLAS data release (Norris et al. 2006) were used to compute accuracy, as these were the only ATLAS objects with expert labels.

We then used the estimated class probabilities to predict the host galaxy for each radio component cross-identified by both Norris et al. (2006) and Radio Galaxy Zoo. For each SWIRE object within 1 arcmin of the radio component, the probability of the object having a positive label was estimated using the trained binary classifiers. The SWIRE object with the highest Gaussian-weighted probability was chosen as the host galaxy. The accuracy for the cross-identification task was then estimated as the fraction of the predicted host galaxies that matched the Norris et al. (2006) cross-identifications.

4 RESULTS

In this section we present accuracies of our method trained on CDFS and applied to CDFS and ELAIS-S1, as well as results motivating our accuracy measures and estimates of upper and lower bounds for cross-identification accuracy using our method.

4.1 Relation between binary classification and cross-identification

We can assess our classifiers either by their performance on the galaxy classification task or by their performance on the cross-identification task. Both performances are useful: performance on the galaxy classification task provides a robust and simple way to compare classifiers without the limitations of our specific formulation, and performance on the cross-identification task can be compared with other cross-identification methods. We therefore report two sets of accuracies: balanced accuracy for the galaxy classification task and accuracy for the cross-identification task. These accuracy measures are correlated and we show this correlation in Figure 10. Fitting a line of best fit with `scipy` gives $R^2 = 0.92$ for logistic regression and $R^2 = 0.87$ for random forests.

While performance on the galaxy classification task is correlated with performance on the cross-identification task, balanced accuracy does not completely capture the effectiveness of a binary classifier applied to the cross-identification task. This is because binary classifiers output a predicted label for each object in addition to the class probability, and it is this predicted label that is used to compute balanced accuracy. In the galaxy classification task, the classifier only needs to maximise the rate at which it correctly identifies host galaxies. This means there can be many 'false positives' that do not affect cross-identification. An example of this is shown in Figure 9, where the classifier has identified 8 'host galaxies'. However, there are only three true host galaxies in this image — one per radio component — and so in the cross-identification task, only three of these galaxies will be identified as hosts.

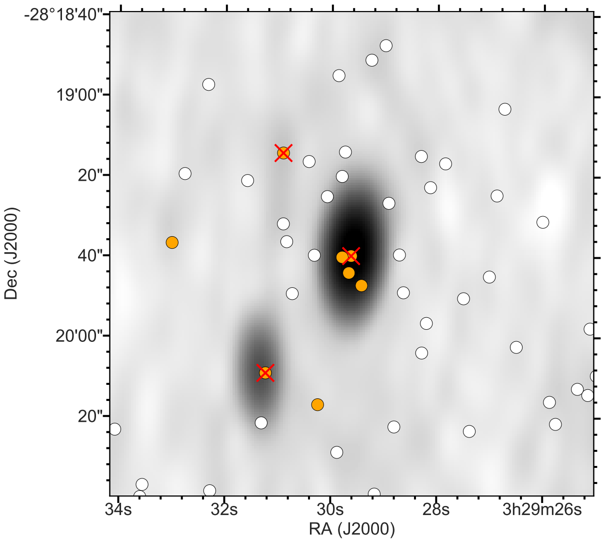


Figure 9. ATLAS3 J032929.61-281938.9. Radio Galaxy Zoo host galaxies are marked by crosses. SWIRE candidate host galaxies are circles, coloured orange circles have been assigned a 'positive' label by a logistic regression classifier and white otherwise.

The formulation of cross-identification as binary classification also introduces limitations as described in section 3.2, which impose upper limits on performance. Additionally, the classification pipeline described in Figure 3 only uses the binary classification approach for non-compact objects, which imposes lower limits on performance. We estimate the upper limit on performance by assuming we have a 'perfect' classifier which simply reads the correct label from the test set, and hence assign all true host galaxies a 100 per cent probability of being host galaxies. The accuracy of this classifier on the cross-identification task then represents the best possible cross-identification performance achievable under our assumptions. We estimate the lower limit on performance by taking a random classifier, which outputs random probabilities regardless of input galaxy. We expect any useful classifier to produce better results than this, so this represents the lowest expected cross-identification performance. We report the cross-identification accuracies of the perfect and random classifiers alongside the performance of other classifiers. We also report the accuracy of a nearest-neighbours approach, where the infrared object nearest to a radio component is chosen as its host galaxy. The upper estimates, lower estimates, and nearest-neighbour accuracy are shown as horizontal lines in Figure 12 for CDFS and Figure 15 for ELAIS-S1.

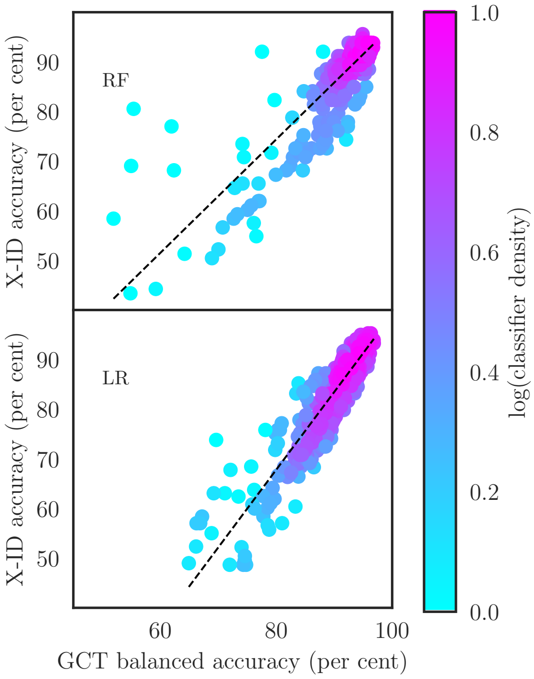


Figure 10. Balanced accuracy on the galaxy classification task (GCT) plotted against accuracy on the cross-identification task (X-ID). RF indicates results from random forests, and LR indicates results from logistic regression. Classifiers were trained on random, small subsets of the training data to artificially restrict their accuracies. Colour shows the density of points on the plot estimated by a Gaussian kernel density estimate. The solid lines indicate the best linear fit; these fits have $R^2 = 0.92$ for logistic regression and $R^2 = 0.87$ for random forests. We did not include convolutional neural networks in this test, as training them is very computationally expensive. These results exclude classifiers with balanced accuracies less than 51 per cent, as these are essentially random.

4.2 Application to ATLAS-CDFS

In Figure 11 we plot the performance of logistic regression, random forests, and convolutional neural networks on both the galaxy classification task and the cross-identification task for resolved and compact sources. For comparison, we also plot the accuracy of Radio Galaxy Zoo and nearest neighbours on the cross-identifiction task. In Figure 12 we plot the performance of these classifiers on the full set of ATLAS DR1 radio components using the pipeline in Figure 3.

Differences between accuracies across training labels are well within error across the four quadrants, with convolutional neural networks on compact objects as the only exception. The spread of accuracies is similar for both sets of training labels, with the exception of random forests. The balanced accuracies of random forests trained on expert labels have a considerably higher spread than those trained on Radio Galaxy Zoo labels, likely because of the small size of the expert training set — there are less than half the number of objects in the expert-labelled training set than the number of objects in the Radio Galaxy Zoo-labelled training set (Table 2).

Radio Galaxy Zoo-trained classifiers significantly outperform

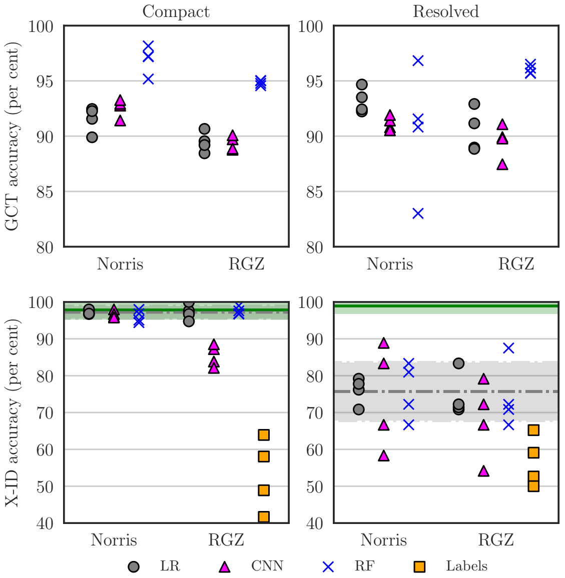


Figure 11. Performance of logistic regression (LR), convolutional neural networks (CNN), and random forest (RF) classifiers. 'Norris' indicates classifiers trained on the expert labels and 'RGZ' indicates classifiers trained on the Radio Galaxy Zoo labels. One point is shown per classifier per testing quadrant. The training and testing sets have been split into compact and resolved objects, with results on compact objects shown on the left and results on resolved objects shown on the right. Shown for comparison is the X-ID accuracy of the Radio Galaxy Zoo consensus cross-identifications (Labels). The X-ID accuracy attained by a perfect classifier is shown by a solid green line, and the X-ID accuracy of a nearest-neighbours approach is shown by a dashed grey line. The standard deviation of these accuracies across the four CDFS quadrants is shown by the shaded area. Note that the pipeline shown in Figure 3 is not used for these results.

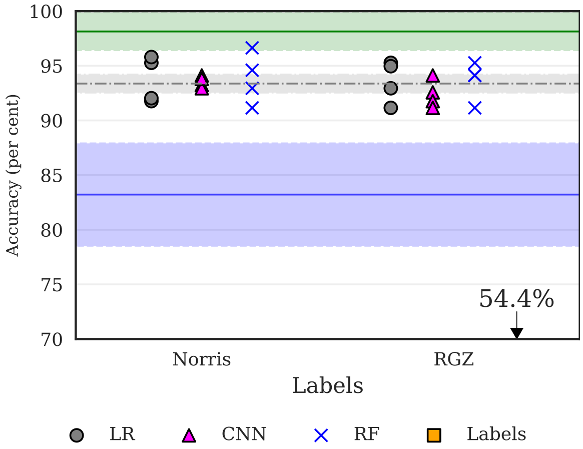


Figure 12. Performance of different classifiers on the cross-identification task. Markers and lines are as in Figure 11. The blue solid line indicates the performance of a random classifier and hence represents the minimum accuracy we expect to obtain. The standard deviation of this accuracy across 25 trials and 4 quadrants is shaded. The Radio Galaxy Zoo accuracies ('Labels') are below the axis and are marked by an arrow with the mean accuracy. Note that the pipeline shown in Figure 3 is used here.

Radio Galaxy Zoo cross-identifications. Additionally, despite poor performance of Radio Galaxy Zoo on the cross-identification task, classifiers trained on these cross-identifications still perform comparably to those trained on expert labels. This is because incorrect Radio Galaxy Zoo cross-identifications can be thought of as a source of noise in the labels which is averaged out in training. This shows the usefulness of crowdsourced training data, even when the data is noisy.

Our classifiers perform comparably to a nearest-neighbours approach. For compact objects, this is to be expected — indeed, nearest neighbours attains nearly 100 per cent accuracy on the compact test set. Our results do not improve on nearest neighbours for resolved objects. However, our method does allow for improvement on nearest neighbours with a sufficiently good binary classifier: a 'perfect' classifier attains nearly 100 per cent accuracy on resolved sources. This shows that our method may be useful provided that a good binary classifier can be trained. The most obvious place for improvement is in feature selection; we use pixels of radio images directly and these are likely not conducive to good performance on the galaxy classification task. Convolutional neural networks, which are able to extract features from images, *should* work better, but these require far more training data than the other methods we have applied and the small size of ATLAS thus results in poor performance.

We noted in section 3.5 that the test set of expert labels, derived from the initial ATLAS data release, was less deep than the third data release used by Radio Galaxy Zoo and this paper, introducing a source of label noise in the testing labels. Specifically, true host galaxies may be misidentified as non-host galaxies if the associated radio source was below the 5 signal-to-noise limit in ATLAS DR1 but not in ATLAS DR3. This has the effect of reducing the accuracy for Radio Galaxy Zoo-trained classifiers.

We report the probabilities predicted by each classifier for each SWIRE object in the supplement. Probabilities reported for a given object were predicted by classifiers tested on the quadrant containing that object.

For the cross-identification task, we report the accuracy of each classification model and training label set in Figure 12. We report the averaged cross-identification accuracies across all four quadrants in the supplement.

The predicted cross-identification for each ATLAS object is reported in the supplement. As with SWIRE predicted probabilities, reported cross-identifications for a given object were generated by classifiers tested on the quadrant containing that object.

In Figure 13 we show 6 of the 16 resolved sources where the incorrect cross-identification was selected by random forests trained on expert labels. The remaining 10 sources are shown in the supplement. In most cases the incorrect galaxy selected was also selected by nearest neighbours. We note that on no source do all classifiers agree. This raises the possibility of using the level of disagreement of an ensemble of classifiers as a measure of the complexity of a radio source, analogous to the consensus level for Radio Galaxy Zoo volunteers.

4.3 Application to ATLAS-ELAIS-S1

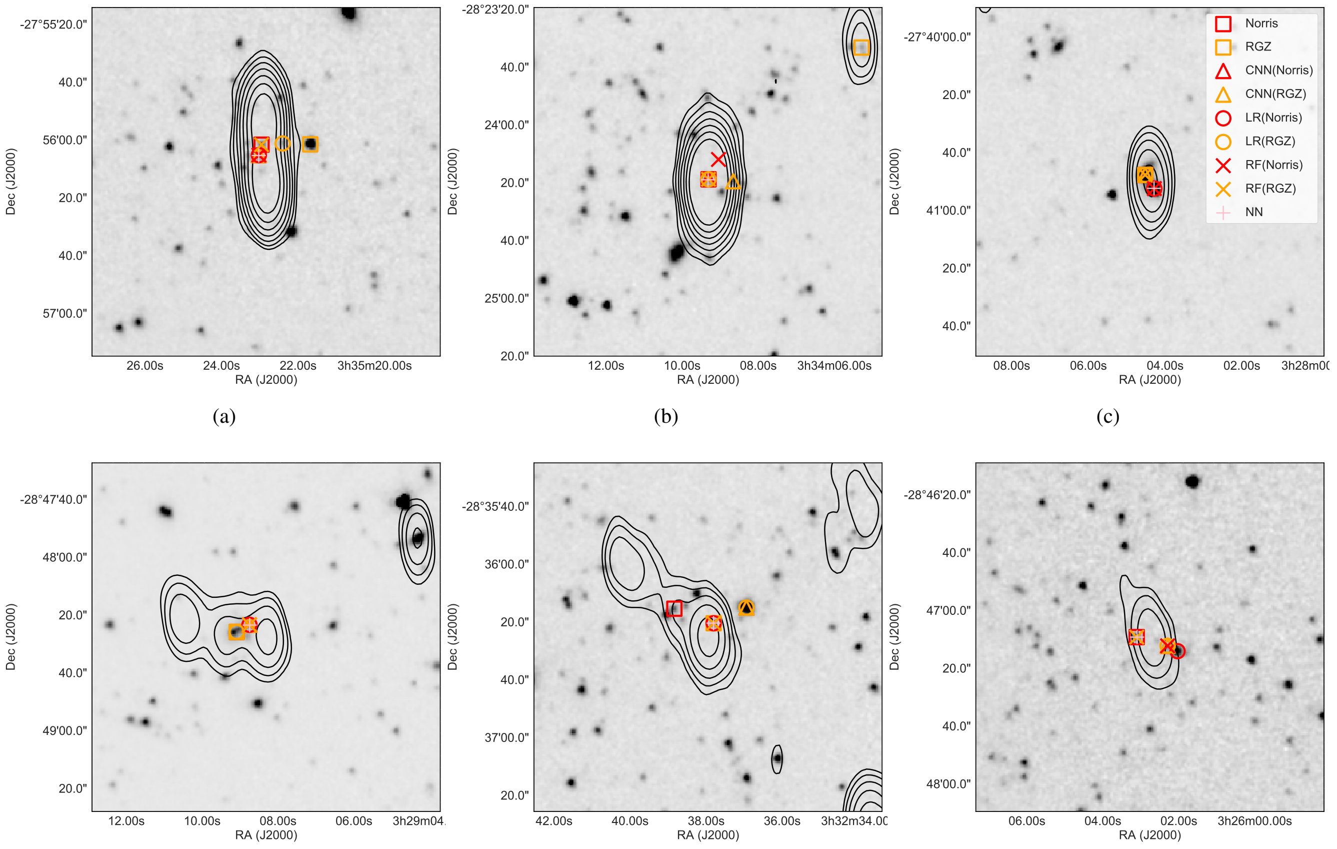
We applied the classifiers trained on CDFS to perform cross-identification on the ELAIS-S1 field. Both CDFS and ELAIS-S1 were imaged by the same radio telescope to similar sensitivities and angular resolution for the ATLAS survey. As ELAIS-S1 has been cross-identified with SWIRE host galaxies by Middelberg et al. (2008), we can use these cross-identifications to derive another set of expert labels, and hence determine how accurate our method is. If our method generalises well across different parts of the sky, then we expect CDFS-trained classifiers applied to cross-identification in ELAIS-S1 to perform comparably to application to CDFS. In Figure 14 we plot the performance of CDFS-trained logistic regression, random forests, and convolutional neural networks on both the galaxy classification task and the cross-identification task for resolved and compact sources. We also plot the accuracy of a nearest-neighbours approach. In Figure 15 we plot the performance of these classifiers on the full set of ATLAS DR1 radio components using the pipeline in Figure 3. We list the corresponding accuracies in the supplement.

Cross-identification results from ELAIS-S1 are similar to those for CDFS, showing that classifiers trained on CDFS perform reasonably well on ELAIS-S1. However, nearest neighbours outperforms most methods on ELAIS-S1. This is likely because there is a much higher percentage of compact objects in ELAIS-S1 than in CDFS. The maximum achievable accuracy we have estimated for ELAIS-S1 is very close to 100 per cent, so (as for CDFS) a very accurate binary classifier would outperform nearest neighbours.

One interesting difference between the ATLAS fields is that random forests trained on expert labels perform well on CDFS but poorly on ELAIS-S1. This is not the case for logistic regression or convolutional neural networks trained on expert labels, nor is it the case for random forests trained on Radio Galaxy Zoo. We hypothesise that this is because the ELAIS-S1 cross-identification catalogue (Middelberg et al. 2008) labelled fainter radio components than the CDFS cross-identification catalogue (Norris et al. 2006) due to noise from the very bright source ATCDFS_J032836.53-284156.0 in CDFS. Classifiers trained on CDFS expert labels may thus be biased toward brighter radio components compared to ELAIS-S1. Radio Galaxy Zoo uses the third data release of ATLAS (Franzen et al. 2015) and so classifiers trained on the Radio Galaxy Zoo labels may be less biased toward brighter sources compared to those trained on the expert labels. To test this hypothesis we tested each classifier against test sets with a signal-to-noise ratio (SNR) cutoff. A SWIRE object was only included in the test set for a given cutoff if it was located within $1'$ of a radio component with a SNR above the cutoff. The balanced accuracies for each classifier at each cutoff are shown in Figure 16(a) and (b) and the distribution of test set size for each cutoff is shown in Figure 16(c). Figure 16(c) shows that ELAIS-S1 indeed has more faint objects than CDFS, with the SNR for which the two fields reach the same test set size (approximately 34) indicated by the dashed vertical line on each plot. For CDFS, all classifiers perform reasonably well across cutoffs, with performance dropping as the size of the test set becomes small. For ELAIS-S1, logistic regression and convolutional neural networks perform comparably across all SNR cutoffs, but random forests do not: while random forests trained on Radio Galaxy Zoo labels perform comparably to other classifiers across all SNR cutoffs, random forests trained on expert labels show a considerable drop in performance below the dashed line.

5 DISCUSSION

Our main result is that it is possible to cast radio host galaxy cross-identification as a machine learning task for which standard methods can be applied. These methods can then be trained with a variety of label sets derived from cross-identification catalogues. While we have not outperformed nearest neighbours, we have demonstrated that for a very accurate binary classifier, good



(a)

(b)

(c)

(d)

(e)

(f)

Figure 13. Examples of resolved sources where the incorrect cross-identification was selected by random forests trained on expert labels. Classifier/training set combinations are denoted $C(S)$ where C is the classifier and S is the training set. ‘LR’ is logistic regression, ‘CNN’ is convolutional neural networks, and ‘RF’ is random forests. ‘Norris’ refers to the expert labels and ‘RGZ’ refers to the Radio Galaxy Zoo labels. The cross-identification made by nearest neighbours is shown by ‘NN’. The background image is the 3.6 μm SWIRE image. The contours show ATLAS radio data and increase geometrically from a signal-to-noise ratio of 4.

cross-identification results can be obtained using our method. Future work could even combine multiple catalogues or physical priors to boost performance.

Nearest neighbours approaches outperform most methods we investigated, notably including Radio Galaxy Zoo. This is due to the large number of compact or partially-resolved objects in ATLAS. This result shows that for compact and partially-resolved objects methods that do not use machine learning such as a nearest-neighbours approach or likelihood ratio (Weston et al. in press) should be preferred to machine learning methods. It also shows that ATLAS is not an ideal data set for developing machine learning methods like ours. Our use of ATLAS is motivated by its status as a pilot survey for EMU, so methods developed for ATLAS should also work for EMU. New methods developed should work well with extended radio sources, but this goal is almost unsupported by ATLAS as it has very few examples of such sources. This makes both training and testing difficult — there are too few extended sources to train on and performance on such a small test set may be unreliable. Larger data sets with many extended sources like FIRST exist, but these are considerably less deep than and at a different resolution to EMU, so there is no reason to expect methods trained on such data sets to be applicable to EMU.

The accuracies of our trained cross-identification methods generally fall far below the best possible accuracy attainable using our approach, indicated by the green lines in Figures 12 and 15. The balanced accuracies attained by our binary classifiers indicate that there is significant room for improvement in classification. The classification accuracy can be improved by better model selection and more training data, particularly for convolutional neural networks. There is a huge variety of ways to build a convolutional neural network, and we have only investigated one architecture. For an exploration of different convolutional neural network architectures applied to radio astronomy, see Lukic et al. (in prep). Convolutional neural networks generally require more training data than other machine learning models and we have only trained our networks on a few hundred sources. We would thus expect performance on the classification task to greatly increase with larger training sets.

Another problem is that of the window size used to select radio features. Increasing window size would increase computational expense, but provide more information to the models. Results are also highly sensitive to how large the window size is compared to the size of the radio galaxy we are trying to cross-identify, with large angular sizes requiring large window sizes to ensure that the features contain all the information needed to localise the host galaxy. An ideal implementation of our method would most likely represent a galaxy using radio images taken at multiple window sizes, but this is considerably more expensive.

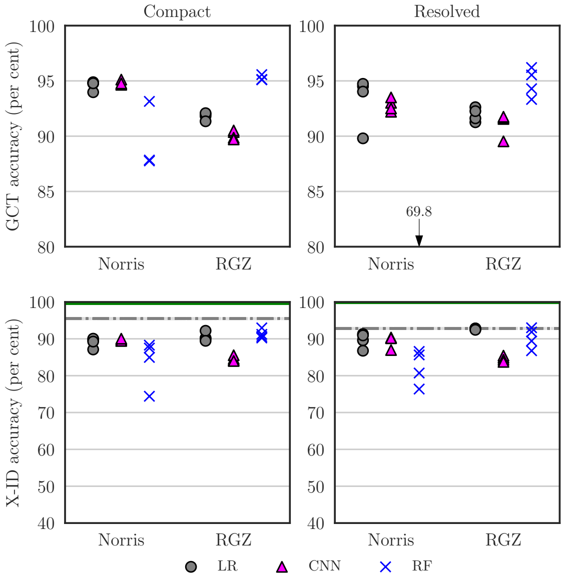


Figure 14. Performance of different classifiers trained on CDFS and tested on resolved and compact sources in ELAIS-S1. Points represent classifiers trained on different quadrants of CDFS, with markers and axes as in Figure 11. The balanced acccuracy of expert-trained random forest classifiers falls below the axis and the corresponding mean accuracy is shown by an arrow. The solid green line indicates the X-ID performance of a 'perfect' classifier; this is almost 100 per cent. The grey dashed line indicates the X-ID performance of nearest neighbours. Note that the pipeline in Figure 3 is not used in this figure.

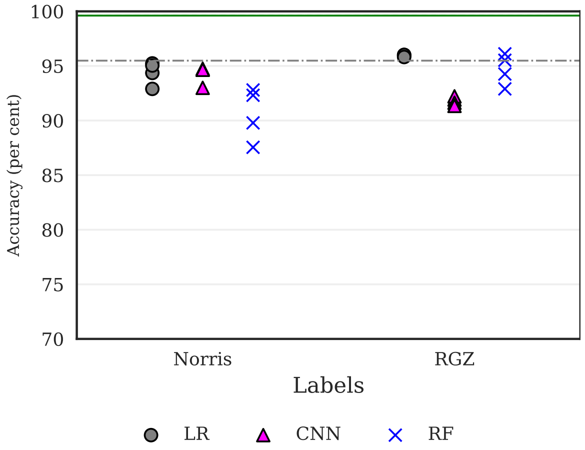


Figure 15. Performance of different classifiers trained on CDFS and tested on ELAIS-S1. Markers are as in Figure 12 and horizontal lines are as in Figure 14. Note that the pipeline in Figure 3 is not used in this figure.

Larger training sets, better model selection, and larger window sizes would improve performance, but only so far: we would still be bounded above by the estimated "perfect" classifier accuracy. From this point, the performance can only be improved by improving upon our broken assumptions. We detailed these assumptions in section 3.2, and we will discuss here how these could be resolved. Our assumption that the host galaxy is contained within the search radius could be improved by dynamically choosing the search ra-

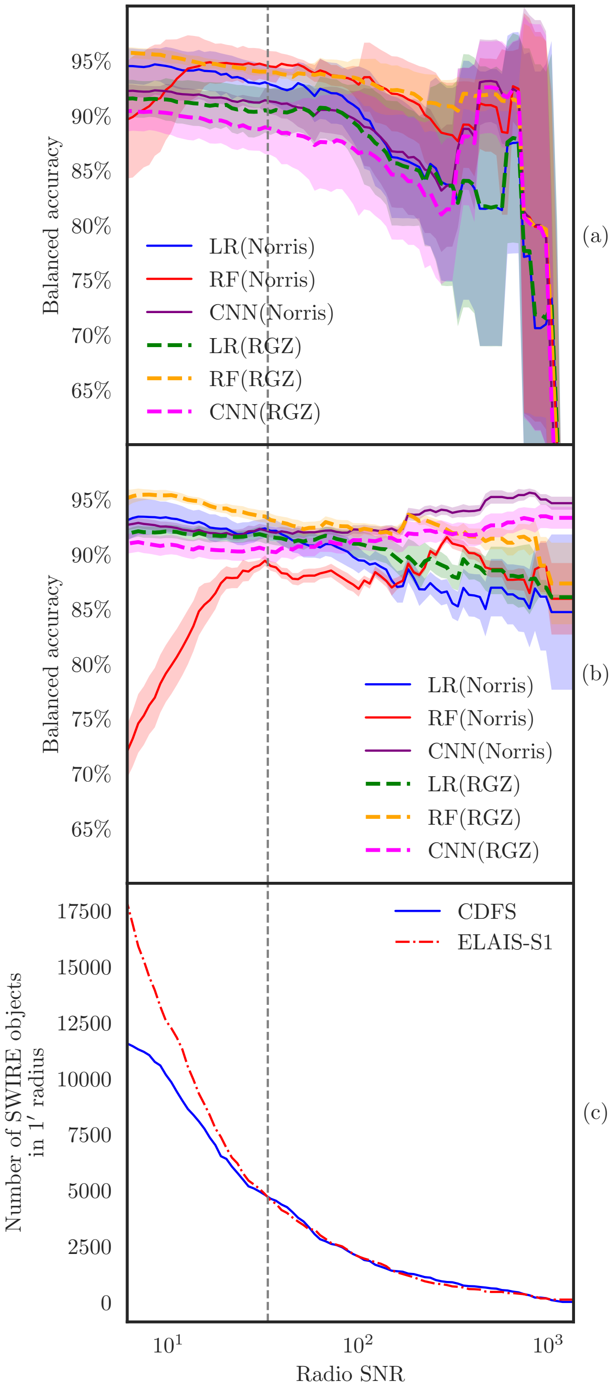


Figure 16. (a) Balanced accuracies of classifiers trained and tested on CDFS with different signal-to-noise ratio (SNR) cutoffs for the test set. A SWIRE object is included in the test set if it is within 1′ of a radio component with greater SNR than the cutoff. Different coloured lines indicate different classifier/training labels combinations, where LR is logistic regression, RF is random forests, CNN is convolutional neural networks, and Norris and RGZ are the expert and Radio Galaxy Zoo label sets respectively. Filled areas represent standard deviations across CDFS quadrants. (b) Balanced accuracies of classifiers trained on CDFS and tested on ELAIS-S1. (c) A cumulative distribution plot of SWIRE objects associated with a radio object with greater SNR than the cutoff. The grey dashed line shows the SNR level at which the number of SWIRE objects above the cutoff is equal for CDFS and ELAIS-S1. This cutoff level is approximately 34 times the RMS.

dius, perhaps based on the angular extent of the galaxy, or the redshift of candidate hosts. Radio morphology information may allow us to select relevant radio data and hence relax the assumption that a $1'$-wide radio image represents just one, whole radio source. Finally, our assumption that the host galaxy is visible in infrared is technically not needed, as the sliding window approach we have employed will still work even if there are no visible host galaxies — instead of classifying candidate hosts, simply classify each pixel in the radio image. The downside of removing candidate hosts is that we are no longer able to reliably incorporate host galaxy information such as colour and redshift, though this could be resolved by treating pixels as potentially invisible candidate hosts with noisy features.

We observe that Radio Galaxy Zoo-trained methods perform comparably to methods trained on expert labels. This shows that the crowdsourced labels from Radio Galaxy Zoo will indeed provide a valuable source of training data for future machine learning methods in radio astronomy.

Compared to nearest neighbours, cross-identification accuracy on ELAIS-S1 is lower than on CDFS. Particularly notable is that our performance on compact objects is very low for ELAIS-S1, while it was near-optimal for CDFS. These differences may be for a number of reasons. ELAIS-S1 has beam size and noise profile different from CDFS (even though both were imaged with the same telescope), so it is possible that our methods over-adapted to the beam and noise of CDFS. Additionally, CDFS contains a very bright source which may have caused artefacts throughout the field that are not present in ELAIS-S1. Further work is required to understand the differences between the fields and their effect on performance.

Figure 16 reveals interesting behaviour of different classifier models at different flux cutoffs. Logistic regression and convolutional neural networks seem relatively independent of flux, with these models performing well on the fainter ELAIS-S1 sources even when they were trained on the generally brighter objects in CDFS. Conversely, random forests were sensitive to the changes in flux distribution between datasets. This shows that not all models behave similarly on radio data, and it is therefore important to investigate multiple models when developing machine learning methods for radio astronomy.

Our methods can be easily incorporated into other cross-identification methods or used as an extra data source for source identification. This is because we have essentially produced a scoring function, which rates galaxies based on the probability that they are a host galaxy. For example, our method could be used to disambiguate between candidate host galaxies selected by model-based algorithms, or used to weight candidate host galaxies while a source identifier attempts to associate radio components. Our method can also be extended using other data sources; for example, information from source identification algorithms could be incorporated into the feature set of candidate host galaxies.

6 SUMMARY

We presented a machine learning approach for cross-identification of radio components with their corresponding infrared host galaxy. Using the CDFS field of ATLAS as a training set we trained our methods on expert and crowdsourced cross-identification catalogues. Applying these methods on both fields of ATLAS, we found that:

- Our method trained on ATLAS observations of CDFS generalised to ATLAS observations of ELAIS-S1, demonstrating that training on a single patch of sky is a feasible option for training machine learning methods for wide-area radio surveys;
- Performance was comparable to nearest neighbours even on resolved sources, showing that nearest neighbours is useful for largely unresolved datasets such as ATLAS and EMU;
- Radio Galaxy Zoo-trained models performed comparably to expert-trained models, showing that crowdsourced labels are useful for training machine learning methods for cross-identification even when these labels are noisy;
- ATLAS does not contain sufficient data to train or test machine learning cross-identification methods for extended radio sources. This suggests that if machine learning methods are to be used on EMU, a larger area of sky will be required for training and testing these methods. However, existing surveys like FIRST are too different from EMU to expect good generalisation.

While our cross-identification performance is not as high as desired, we make no assumptions on the binary classification model used in our methods and so we expect the performance to be improved by further experimentation and model selection. Our method provides a useful framework for generalising cross-identification catalogues to other areas of the sky from the same radio survey and can be incorporated into existing methods.

7 ACKNOWLEDGEMENTS

This publication has been made possible by the participation of more than 11 000 volunteers in the Radio Galaxy Zoo project. Their contributions are individually acknowledged at http://rgzauthors.galaxyzoo.org. Parts of this research were conducted by the Australian Research Council Centre of Excellence for All-sky Astrophysics (CAASTRO), through project number CE110001020. Partial support for LR was provided by U.S. National Science Foundation grants AST1211595 and 1714205 to the University of Minnesota. HA benefitted from grant 980/2016-2017 of Universidad de Guanajuato. We thank A. Tran for his comments on this manuscript. Radio Galaxy Zoo makes use of data products from the Wide-field Infrared Survey Explorer and the Very Large Array. The Wide-field Infrared Survey Explorer is a joint project of the University of California, Los Angeles, and the Jet Propulsion Laboratory/California Institute of Technology, funded by the National Aeronautics and Space Administration. The National Radio Astronomy Observatory is a facility of the National Science Foundation operated under cooperative agreement by Associated Universities, Inc. This research has made use of the NASA/IPAC Extragalactic Database (NED) which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration. The figures in this work made use of Astropy, a community-developed core Python package for Astronomy (Astropy Collaboration et al. 2013). The Australia Telescope Compact Array is part of the Australia Telescope, which is funded by the Commonwealth of Australia for operation as a National Facility managed by CSIRO.

References

Aniyan A. K., Thorat K., 2017, ApJS, 230, 20
Astropy Collaboration et al., 2013, A&A, 558, A33
Banfield J. K., et al., 2015, MNRAS, 453, 2326
Bishop C. M., 2006, Pattern recognition and machine learning. Springer

Breiman L., 2001, Machine Learning, 45, 5
Dieleman S., Willett K. W., Dambre J., 2015, MNRAS, 450, 1441
Fan D., Budavri T., Norris R. P., Hopkins A. M., 2015, MNRAS, 451, 1299
Fanaroff B. L., Riley J. M., 1974, MNRAS, 167, 31P
Franzen T. M. O., et al., 2015, MNRAS, 453, 4020
Gendre M. A., Wall J. V., 2008, MNRAS, 390, 819
Grant J. K., Taylor A. R., Stil J. M., Landecker T. L., Kothes R., Ransom R. R., Scott D., 2010, ApJ, 714, 1689
Johnston S., et al., 2007, PASA, 24, 174
Lintott C. J., et al., 2008, MNRAS, 389, 1179
Lonsdale C. J., et al., 2003, PASP, 115, 897
Middelberg E., et al., 2008, AJ, 135, 1276
Norris R. P., 2017, PASA, 34, e007
Norris R. P., et al., 2006, AJ, 132, 2409
Norris R. P., et al., 2011, PASA, 28, 215
Proctor D. D., 2006, ApJS, 165, 95
Richter G. A., 1975, Astronomische Nachrichten, 296, 65
Rowley H. A., Baluja S., Kanade T., 1996, in Advances in Neural Information Processing Systems. pp 875–881
Sajina A., Lacy M., Scott D., 2005, ApJ, 621, 256
Surace J., et al., 2005, Spitzer Science Centre, California Institute of Technology, Pasadena, CA
Taylor A. R., et al., 2007, ApJ, 666, 201
Verheijen M. A. W., Oosterloo T. A., van Cappellen W. A., Bakker L., Ivashina M. V., van der Hulst J. M., 2008, in AIP Conf. Ser. 1035, The Evolution of Galaxies Through the Neutral Hydrogen Window. pp 265–271
White R. L., Becker R. H., Helfand D. J., Gregg M. D., 1997, ApJ, 475, 479

This paper has been typeset from a TEX/LATEX file prepared by the author.