Machine learning methods for radio host cross-identification with crowdsourced labels

Matthew Alger[1] matthew.alger@anu.edu.au

Julie Banfield[1,2,3] julie.banfield@anu.edu.au

Cheng Soon Ong[4,5] cheng-soon.ong@data61.csiro.au

Ivy Wong[2,6] ivy.wong@uwa.edu.au

and the Radio Galaxy Zoo team

[1]Research School of Astronomy and Astrophysics, The Australian National University, Canberra, Australia
[2]ARC Centre of Excellence for All-Sky Astrophysics (CAASTRO)
[3]Western Sydney University, Penrith, Australia
[4]Data61, CSIRO, Canberra, Australia
[5]Research School of Computer Science, The Australian National University, Canberra, Australia
[6]International Centre for Radio Astronomy Research, University of Western Australia, Perth, Australia

Outline

We propose a machine learning approach for radio source host galaxy cross-identification, the task of finding the infrared host galaxy corresponding to radio emissions. We train our method on both expert cross-identifications and Radio Galaxy Zoo cross-identifications of the Australia Telescope Large Area Survey (ATLAS; Norris et al. 2006) survey of the *Chandra* Deep Field – South (CDFS). Radio Galaxy Zoo (RGZ; Banfield et al. 2015) is a citizen science project that allows volunteers to cross-identify radio emissions with their infrared host galaxies. These non-expert cross-identifications may be incorrect, but are nevertheless useful for training, and we find that machine learning methods trained on the Radio Galaxy Zoo cross-identifications perform comparably to those trained on expert cross-identifications.

Task

- Goal: Identify the infrared host galaxy of radio emission in an infrared and radio image.
- Approach: Machine learning model trained on Radio Galaxy Zoo host cross-identifications.
- Use:
  - Future radio survey processing pipelines.
  - Case study into benefits and problems associated with applying machine learning to radio astronomy.

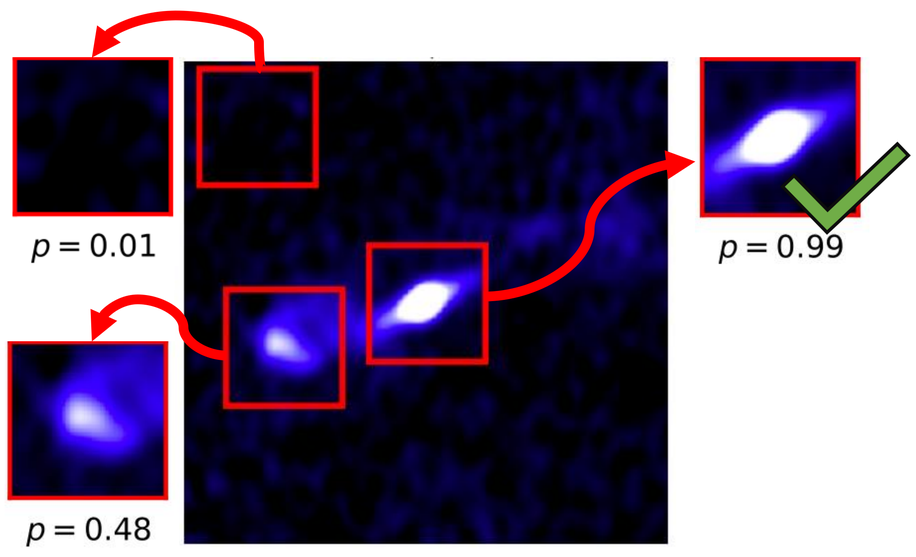
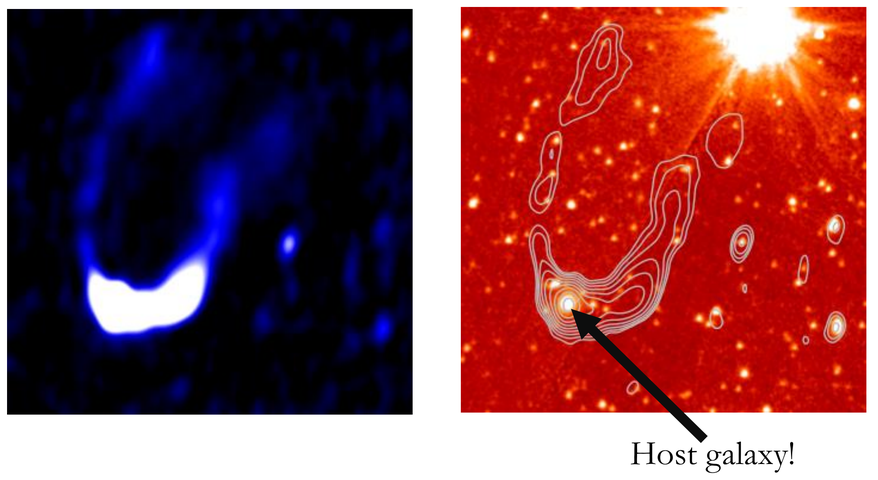


Figure 1: *Predicting the probability that candidate locations coincide with the host galaxy. The highest probability candidate location is selected as the host galaxy location.*



Approach

- Convert cross-identification into **binary classification**, a well-understood machine learning problem.
- Classify all nearby infrared galaxies as either being the host galaxy or not.
  - Each galaxy is represented by a 32 × 32 pixel radio image centred on that location.
- Select the galaxy with the highest probability of being the host galaxy (**Figure 1**).

Experimental Results

- We trained three classification models on two datasets, one derived from expert cross-identifications (Norris et al. 2006) and one derived from Radio Galaxy Zoo.
  - Compact objects were used for training, but were cross-identified separately using a nearest-neighbour approach.
- Expert-trained models outperformed those trained on Radio Galaxy Zoo, but Radio Galaxy Zoo-trained models still performed well (**Figure 2**).
- Problems:
  - Method is sensitive to search radius and width of the radio image representing each galaxy.
  - Too small search radius might miss host galaxy, too large might include multiple.
  - Deep radio surveys are more sensitive to these parameters when compared to shallower surveys due to the increased density of radio objects.
  - Radio morphology information could resolve this.

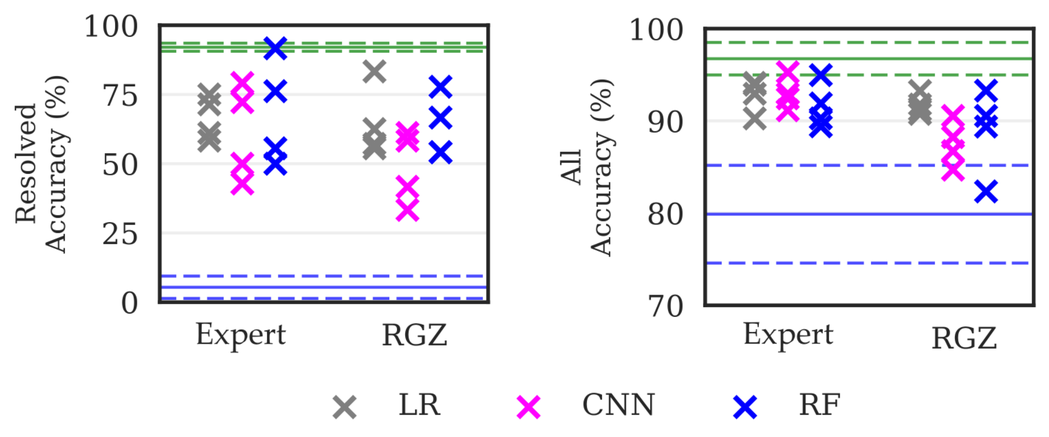
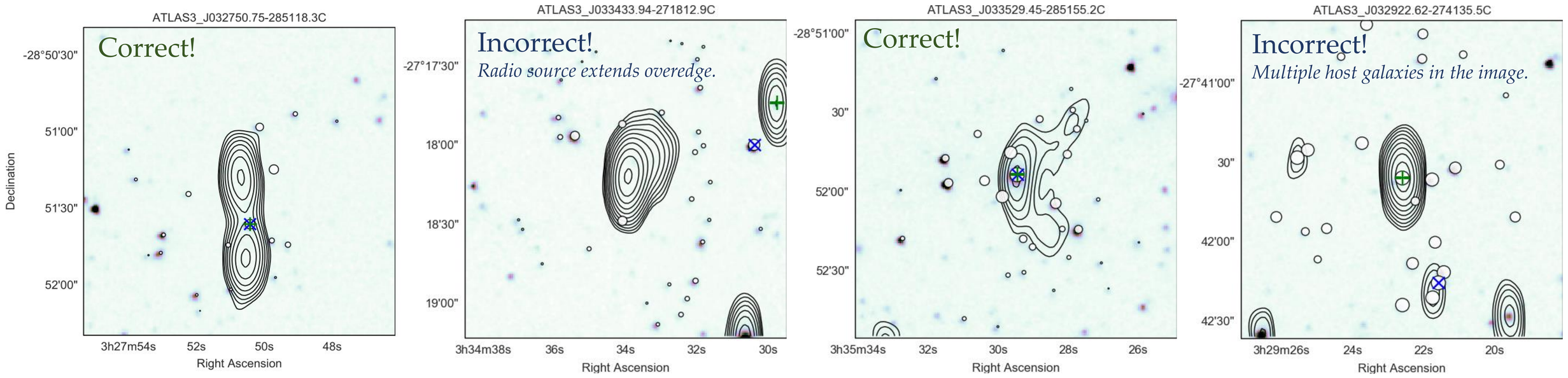


Figure 2: *Cross-identification accuracy on four subsets of CDFS. We trained three binary classification models: logistic regression (LR), convolutional neural networks (CNN), and random forests (RF). We used two training sets, one derived from expert cross-identifications (Norris et al. 2006) and one derived from Radio Galaxy Zoo (Banfield et al. 2015). The solid green line indicates an estimate of the best possible accuracy attainable under our assumptions, and the blue solid line indicates an estimate of the minimum accuracy with random classifiers. Dashed lines indicate standard deviation over quadrants and classifiers. "All" includes both compact and resolved radio objects.*



× *Predicted host galaxy* + *True host galaxy (from Norris et al. 2006)*

Example cross-identifications performed by our convolutional neural network.. The background image is an infrared image from Spitzer, *and the contours are radio from ATLAS. Circles indicate candidate host galaxies, and circle size is proportional to predicted probability that the candidate is a host galaxy. The predicted and true host galaxies are marked.*

References: Banfield J. K., et al., 2015, MNRAS, 453, 2326; Norris R. P., et al., 2006, AJ, 132, 2409

Acknowledgements: Data for this project comes from the ATCA and the Spitzer Space Telescope. SWIRE is supported by NASA through the SIRTF Legacy Program under contract 1407 with the Jet Propulsion Laboratory. The ATCA is part of the Australia Telescope, which is funded by the Commonwealth of Australia for operation as a National Facility managed by CSIRO. This work has been made possible by the participation of over 10,000 volunteers in the Radio Galaxy Zoo project. Their contributions are individually acknowledged at http://rgzauthors.galaxyzoo.org.