Transfer learning for radio galaxy classification

H. Tang [ORCID],⋆ A. M. M. Scaife and J. P. Leahy

Jodrell Bank Centre for Astrophysics, University of Manchester, Oxford Road, Manchester M13 9PL, UK

ABSTRACT

In the context of radio galaxy classification, most state-of-the-art neural network algorithms have been focused on single survey data. The question of whether these trained algorithms have cross-survey identification ability or can be adapted to develop classification networks for future surveys is still unclear. One possible solution to address this issue is *transfer learning*, which reuses elements of existing machine learning models for different applications. Here we present radio galaxy classification based on a 13-layer Deep Convolutional Neural Network (DCNN) using transfer learning methods between different radio surveys. We find that our machine learning models trained from a random initialization achieve accuracies comparable to those found elsewhere in the literature. When using transfer learning methods, we find that inheriting model weights pre-trained on FIRST images can boost model performance when retraining on lower resolution NVSS data, but that inheriting pre-trained model weights from NVSS and retraining on FIRST data impairs the performance of the classifier. We consider the implication of these results in the context of future radio surveys planned for next-generation radio telescopes such as ASKAP, MeerKAT, and SKA1-MID.

Key words: methods: statistical – surveys – radio continuum: galaxies.

1 INTRODUCTION

The radio morphology of Double Radio sources associated with Active Galactic Nuclei (DRAGNs; Leahy 1993) is typically determined by examining the distribution of high and low surface brightness regions in the radio synchrotron emitting relativistic jets associated with these systems. This morphology is the basis of the well-known Fanaroff–Riley radio galaxy classification scheme (FR; Fanaroff & Riley 1974). This classification splits radio galaxies into two groups, known as FR I and FR II, respectively. Sources where the ratio of the distance between the highest brightness regions (known as hotspots) to the total extent of the source is smaller than 0.5. are classified as FR I. Radio galaxies with larger distance ratios are identified as FR II sources (Fanaroff & Riley 1974). The FR classification scheme is also known to be related to a number of other morphological characteristics (Leahy 1993) including jet opening angle (Bridle & Perley 1984; Perucho et al. 2014). Moreover, although FR classification is primarily made on source morphology it is also strongly correlated with radio luminosity (Fanaroff & Riley 1974).

Traditionally, identifying FR class is done by visual inspection and such an approach has been widely used over the past few decades (Cotter, Rawlings & Saunders 1996; Leahy, Bridle & Strom 1996; Subrahmanyan, Saripalli & Hunstead 1996; Ishwara-Chandra & Saikia 1999; Lara et al. 2001; Machalski, Jamrozy &

Zola 2001; Schoenmakers et al. 2001; Saripalli et al. 2005; Machalski & Jamrozy 2006; Machalski et al. 2006; Machalski, Koziel-Wierzbowska & Jamrozy 2007; Solovyov & Verkhodanov 2011; Kuźmicz & Jamrozy 2012; Butenko & Tyul'bashev 2016; Dabhade et al. 2017). However, we note that this approach is constrained by the resolution and sensitivity of given radio data. With poor resolution, FR IIs with extended structure may be only partially resolved, with no clear hotspots, and thus be misclassified as FR Is.

A visual inspection method has been widely used in part due to radio survey sample sizes. For instance, the NRAO VLA Sky Survey (NVSS; Condon et al. 1998), Sydney University Molonglo Sky Survey (SUMSS; Bock, Large & Sadler 1999), and the Faint Images of the Radio Sky at Twenty-Centimetres (FIRST; Becker, White & Helfand 1995) have catalogued sources of no more than 2 million objects, respectively (White et al. 1997; Condon et al. 1998; Mauch et al. 2003). Although still large, the size of these catalogues enabled the possibility of finishing the work above by visual inspection. Such sample sizes also make citizen science projects like Radio Galaxy Zoo (RGZ; Banfield et al. 2015) possible.

Radio catalogues produced by the next generation of radio surveys, however, are anticipated to be much larger. The Australia SKA Pathfinder (ASKAP; Johnston et al. 2008) Evolutionary Map of the Universe (EMU; Norris et al. 2011) survey is expected to produce a catalogue of ∼70 million sources. Among the objects in this catalogue, around 7 million extended radio sources are likely to require visual inspection.

⋆ E-mail: hongming.tang@manchester.ac.uk

Beyond ASKAP, the Square Kilometre Array (SKA; Carilli & Rawlings 2004) will be able to observe FR Is, low- and high-luminosity FR IIs up to redshift $z \sim 4$ (Kapinska et al. 2015). This telescope array will have better resolution and sensitivity than current best radio telescopes (McAlpine et al. 2015) and thus should also discover a huge number of extended radio sources.

The anticipated volume of objects from new radio surveys have motivated the introduction of semi-automatic and automatic object classification algorithms. Recently, convolutional neural networks (CNNs; Krizhevsky, Sutskever & Hinton 2012) were applied to radio galaxy morphology classification (Aniyan & Thorat 2017; Ma et al. 2018). These studies suggest that complex radio source structures can be identified and classified according to their morphology from images drawn from a single survey.

However, although such machine learning approaches are useful, training a model often requires a considerable amount of labelled data. One practical way to collect these labels is through citizen science. RGZ, a representative astronomical citizen science project, is primarily used to cross-identify infrared hosts and their corresponding radio lobes. The project has ∼12 000 users and has already made over 2 million classifications since it launched. These classifications have contributed to the foundation of the RGZ DR1 data set (RGZ Data Release 1; Wong et al., in preparation). The data set contains over 70 000 candidate sources, with source coordinates, fluxes, angular extent recorded. The project team have now developed several machine learning algorithms for doing infrared-radio source cross-identification, compact-resolved source separation, radio galaxy localization, and peak-component morphology classification with the data set (Alger et al. 2018; Lukic et al. 2018; Wu et al. 2019). In addition, efforts have been made to reduce the sample numbers requiring visual inspection using unsupervised learning methods such as self-organizing maps (Kohenen 1982). This semi-automatic approach retrieves representative images that share morphological similarity, assisting astronomers to perform more efficient morphological analysis on pre-analysed data (Polsterer, Gieseke & Igel 2015).

Regardless of data collection methodology, it is still unclear whether the project outcomes for a single survey are transferable to other survey data. How to build a data set cost efficiently and maximize the generalization of machine learning models across multiple surveys remain open questions. One solution to these two questions, however, might be transfer learning (Yosinski et al. 2014).

In this paper, we investigate the applicability of transfer learning to the classification of radio galaxy morphology using survey data. In Section 2.1 we introduce the use of CNNs for classification and in Section 2.2 we describe the practice of transfer learning. Construction of our training, test, and validation data sets, including data acquisition, image pre-processing and data formatting, are described in Section 3. Section 4 covers the network architecture adopted in this study and the transfer learning strategies. In Section 5, we compare and discuss the performance of these strategies, and in Section 6 we discuss the applicability of our results to future radio surveys.

In what follows we assume a ΛCDM cosmology with $\Omega_{\rm m}$ = 0.3153, Ω_{Λ} = 0.6847, and a Hubble constant of H_0 = 67.36 km s^{-1} Mpc^{-1} (Planck Collaboration I 2018). Computational work in this paper was done using a system with 32 Intel Xeon(R) E5-2640 v3 CPUs at 2.6 GHz and 202.5 GB memory.

2 THEORY

2.1 Classification using neural networks

In recent years, CNNs have been widely adopted for a variety of applications in image recognition (e.g. Russakovsky et al. 2015), video analysis (e.g. Huang et al. 2018), and natural language processing (e.g. Grefenstette et al. 2014). For astronomy, CNNs have been used to identify galaxy clusters and filaments (Gheller, Vazza & Bonafede 2018), detect fast radio bursts (Connor & van Leeuwen 2018), localize radio galaxies and identify their morphologies (Aniyan & Thorat 2017; Wu et al. 2019), recognize strong gravitational lenses (Metcalf et al. 2018), and to classify supernovae (Kimura et al. 2017).

CNNs are widely accepted in image classification for several reasons. First, CNNs decompose images into multiple patches that are partially overlapped, where each cortex neuron only corresponds to a single patch (Matsugu et al. 2003). This enables the network to classify images with little data pre-processing or a priori feature value designation (Krizhevsky et al. 2012). Secondly, CNNs are *weight sharing*. With shared weights, CNNs can learn features from a single sample via backpropagation and then share those weights with other samples. Thirdly, as a consequence of weight sharing, convolutional networks are translationally invariant. This last characteristic leads to efficiency as it decreases the required input sample size necessary to train a model with good prediction capability.

CNNs use a number of multilayer perceptrons (MLP; Rosenblatt 1961) to learn both machine learning features and classification weights through a supervized learning method known as backpropagation (Goodfellow, Bengio & Courville 2016). Apart from the first and the final layer of a network, every layer treats outputs from the previous layer as inputs and forwards its own outputs to the following layer. Inputs to the first layer are sample images, potentially with multiple channels (e.g. 1 for a grey-scale image or 3 for an RGB image). Outputs from the final layer are used for class prediction. The layers between the input and output layer are known as *hidden* layers and include convolutional layers, fully connected layers, pooling layers, and loss layers. The functionality of these layers will be described below.

For CNNs, multilayer perceptrons are created by applying non-linear activation functions to every convolutional and fully connected layer output. These activation functions introduce non-linearity to the network (Duda, Hart & Stork 2012). When a network is linear, multilayer perceptrons can be summed up to perform as a single-layer perceptron. A single-layer perceptron produces an output that is a linear combination of the inputs and thus has limited ability for feature recognition and classification. The addition of non-linearity at each layer boosts the expressive ability of a network (Minsky & Papert 1988). Representative activation functions include the sigmoid activation function and the Rectified Linear Unit (ReLU; Hahnloser et al. 2000) function. Defined as `f(y) = max(0,y)`, the ReLU function can help to build sparser models and is therefore beneficial when building deeper training networks (Glorot, Bordes & Bengio 2011). Changpinyo, Sandler & Zhmoginov (2017) claim that, compared to the sigmoid function, models trained with the ReLU activation function have higher efficiency and better overall precision.

CNNs learn via a method called 'backward propagation of errors' (Goodfellow et al. 2016) combined with gradient descent

optimization algorithms (Sebastian 2016). The goal of this learning is to minimize the loss function between observation and prediction (Song et al. 2016; Heas 2018). In the context of categorical classification, a representative loss function called *categorical cross-entropy*, `H(y,q)`, is generally used:

$$H(y, q) = -(y\ln(q) + (1 - y)\ln(1 - q)). \tag{1}$$

For FR binary classification, `y` indicates if class label (0 or 1) is classified as FR I for an observation. `q`, on the other hand, refers to the predicted probability that an observation is of FR I class. Ideally, if the predictions match the observations perfectly, `H(y,q)` will become zero.

Overall, CNN optimization can be summarized as a two-phase circular process of propagation and weight updates. During the propagation phase, input samples propagate through the network and form a network output based on initial network weights and biases. The network then compares the desired model output and the actual network output by measuring their cross-entropy. Calculated errors for the output layer then propagate back through the network. Backpropagation finishes when all neurons in the network have equivalent error values. The backpropagation method then uses these values to calculate the gradient of the cross-entropy loss. The resulting gradient becomes the starting point of the optimization phase. In this second phase, backpropagation passes the gradient to a gradient descent optimization algorithm and updates the network weights.

In practice, a typical convolutional network contains:

(i) **Convolutional layers:** Each convolutional layer uses a number of learnable filters. The specified number of filters is known as the *depth* of the layer. Each filter is applied across the full volume of the input data, iteratively calculating the local dot product of the filter with the input. Often a convolutional layer will be combined with an activation step, which reintroduces non-linearity into the network following the linear convolution operation. A typical activation function that is applied to the output of a convolutional layer is the ReLU function described above.

(ii) **Pooling layers:** Pooling layers provide a form of down sampling. A pooling layer partitions the input from the previous layer into non-overlapping segments and calculates a spatial pooling function for that segment. The function applied is designed to preserve the most important information in each case. Such functions may include returning the maximal value for that segment (max pooling) or the average value (average pooling). Pooling layers reduce the dimensionality of the data within the network and hence the required computation; they also have the effect of reducing overfitting by making the network insensitive to small distortions and variations in the data.

(iii) **Fully connected layers:** Fully connected layers are used to implement classification within the network. Neurons within fully connected layers are connected to all activated outputs in the previous layer. Activations can then compute with learnable filters by matrix computation and a bias offset.

(iv) **Loss layer:** The loss layer usually acts as the final network layer. Its outcome often represents a categorical distribution and evaluates how the network prediction deviates from the truth.

With two convolutional layers, two max-pooling layers, and one fully connected layer, Gheller et al. (2018) were able to generate a classifier to identify candidate galaxy clusters and filament structures from simulated image data. The model correctly classified more than 90 per cent of the unlabelled test images, i.e. the model accuracy was higher than 90 per cent.

In addition to the standard network layers described above, dropout layers (Srivastava et al. 2014) and batch normalization (BN) layers (Ioffe & Szegedy 2015) have also been developed. The purpose of the dropout function is to remove weakly connected neurons and thus avoid overfitting. BN, on the other hand, is designed to speed up the training process and regularize a neural network. If each learnable layer of a model uses either BN or dropout functions, the model can become more computationally efficient especially in the context of image classification (Szegedy et al. 2015; Howard et al. 2017). Meanwhile, a models which simultaneously applied both functions on each learnable layers reported impaired or poor performance (Ioffe & Szegedy 2015; Li et al. 2018). It seems that an appropriate strategy is to choose one out of the two functions for each learnable model layer.

A representative example is a network designed by Krizhevsky et al. (2012). The 12-layer convolutional network applies BN on convolutional layers and dropout on fully connected layers. In the context of an astronomical applications, Aniyan & Thorat (2017) slightly modified this architecture, and their model reached over 90 per cent accuracy when performing three-class radio galaxy morphology (FR I, FR II, and Bent-tailed sources) classification with given test images.

2.2 Transfer learning

Constrained by the lack of large-scale labelled astronomical training data, as well as limited computational power, recent astronomical applications have turned their attention to pre-trained models (Wu et al. 2019). These models were trained with ImageNet, a sizeable visual data set with a thousand object categories and millions of labelled sample images from daily life (Russakovsky et al. 2015). As expected, the performance of these generalized models when applied to astronomical survey data was found to be inferior to that of custom models for direct classification (Lukic et al. 2018).

An alternative to direct classification using inherited weights is to initialize from pre-trained model weights before performing customized training. One can restore model weights for either just the initial (Domínguez Sánchez et al. 2019; Wu et al. 2019) or all layers in a network (Ackermann et al. 2018; Domínguez Sánchez et al. 2019). In some cases, such an approach has been shown to help customized models to avoid overfitting and/or to help them to accelerate the training process (Ackermann et al. 2018; Wu et al. 2019).

Known as *transfer learning*, this approach can reuse knowledge from a solved problem and apply it to relevant but different applications (Pratt 1993). For instance, learning general features from handwritten digits and applying them in handwritten character recognition (Maitra, Bhattacharya & Parui 2015). In the context of classifying FR morphology in radio galaxies with CNNs, generic features learned from initial network layers, such as source edges and hotspots, should exist irrespective of the radio survey; complex features, however, would be learned by the last few layers (Aniyan & Thorat 2017). Using this approach to mitigate against model overfitting during the training process is referred to as regularization (Ackermann et al. 2018).

The application of transfer learning for classification requires careful consideration of which layers to train (Sonntag et al. 2017). When applying transfer learning, a network is trained on a new data set of samples using the stored weights for some or all layers from a pre-trained network as an initialization, rather than initializing weights randomly. The choice of which layers to restore depends on the size of the data set used to train the pre-trained model, the

size of the new data set, the correspondence of the two data sets, and the architecture of the pre-trained network.

Here, we consider the use of pre-trained models trained on radio survey data. Rather than learning everyday object features, layers of the CNN learn radio galaxy features such as jets and hotspot relative positions. These features are universal in classifying source FR morphology, irrespective of survey.

The application of transfer learning has two major advantages. First, by inheriting general morphological features from a pre-trained model, a new model may have a better starting point. Secondly, freezing the weights for the convolutional layers can significantly reduce the training time required to achieve comparable accuracy. However, although loading weights from pre-trained models is often practical, the method needs to be treated carefully. High-level features needed to be trained on customized data sets to learn features for a specific classification problem. In addition, as a pre-trained model and customized model are usually addressing different objectives, transfer learning might produce a `NaN` loss function value (Wu et al. 2019).

Furthermore, transfer learning strategies vary depending on network architecture and data set definitions (Aniyan & Thorat 2017). Inappropriate architectures can lead to overfitting, which can become severe when the retrained model uses only a small number of new training samples but has a considerable number of learnable parameters (Wu et al. 2019). The relative training sample size for pre-trained and transfer-learning models requires careful consideration in order to obtain good model performance (Domínguez Sánchez et al. 2019).

In the context of classifying radio morphology, the influence of transfer learning using data from surveys at different frequencies and with different resolution remains unclear. Here, we explore this question using a transfer learning approach implemented on a variation of the AlexNet CNN (Krizhevsky et al. 2012; Aniyan & Thorat 2017). As part of this work:

(i) We develop a quasi-automatic pipeline to construct training data sets from archival radio surveys. This can be used to download and process images from various surveys. This pipeline makes comparing models trained on different surveys possible.

(ii) We convert the data samples to a data set format consistent with the standard MNIST machine learning data set (Lecun et al. 1998). This enables the data sets to be used in other network architectures.

(iii) We simplify and modify the AlexNet CNN architecture (Krizhevsky et al. 2012), a widely accepted CNN architecture recently adopted in radio galaxy classification (Aniyan & Thorat 2017). The primary network requires parallel GPU computation, which is not considered in this work. Our resulting network can be trained and tested with modest computation power to provide end-to-end training and classification.

(iv) We develop and implement three different transfer learning strategies. Pre-trained models are trained on either NVSS (Condon et al. 1998) or FIRST (Becker et al. 1995) images, with final transfer learning models transfer learned on the other.

(v) We demonstrate that the architecture can be used to classify FR radio morphology and achieve accuracy comparable with the performance of human radio astronomers.

(vi) We evaluate the feasibility, training time cost, and model performance when applying different transfer learning strategies. We examine the possibility of using the same classifier to make prediction on both NVSS and FIRST images.

This work provides an alternative method to full network training for radio galaxy classification and demonstrates under what circumstances such an approach is valid.

3 DATA SAMPLE CONSTRUCTION

The data sample used in this work is designed to train a machine learning model capable of automatically classifying FR I and FR II radio galaxies. We use input data extracted from the NVSS (Condon et al. 1998) and FIRST (Becker et al. 1995) radio surveys. Here we describe the data selection process as well as the automated acquisition methods used.

3.1 Astroquery-based image data batch download

We selected an input sample of radio galaxies following a similar methodology to Aniyan & Thorat (2017). This method uses the Combined NVSS and FIRST Galaxies catalogues (CoNFIG; Gendre & Wall 2008; Gendre, Best & Wall 2010), and the FRICAT catalogue (Capetti, Massaro & Baldi 2017) to select objects from the FIRST surveys. These catalogues were selected as they share significant source populations, the data are free accessible, and sources are well-resolved in the FIRST images (Aniyan & Thorat 2017).

The CoNFIG catalogue contains 859 resolved sources divided into four subsamples (CoNFIG 1, 2, 3, and 4) with flux density limits of $S_{1.4\,\mathrm{GHz}} \geq 1.3, 0.8, 0.2, 0.05$ Jy, respectively. These sources are selected from the NVSS survey within the northern field of the FIRST survey.

Redshifts for the catalogue sources were obtained either from SIMBAD, the NASA Extragalactic Data base (NED) or estimated by using the K_s–z relationship (Gendre et al. 2010) with source K_s magnitudes from the 2MASS survey (Skrutskie et al. 2006) where available. From the full sample, 638 sources were associated with redshifts.

Source morphologies in the CoNFIG catalogue were manually identified by looking at their NVSS and FIRST contour maps. Objects were classified as FR I, FR II, *compact*, or *uncertain* (Gendre et al. 2010). Sources with collimated jets, showing hotspots close to their cores, were classified as FR I; sources with their lobes aligned and hotspots situated at the edges of the lobes were classified as FR II. If the source morphology was ambiguous, it was classified as *uncertain*. Sources with sizes smaller than 1 arcsec were classified as *compact*. In all, 95.7 per cent of CoNFIG sources have their radio morphology classified. We note that the FR I sample in this catalogue also includes sources that are considered to be wide-angle tail or irregular by other studies (Leahy 1993).

Depending on whether a candidate in the CoNFIG sample showed a clear FR I or FR II morphology, the identification of each object was qualified as 'confirmed' (c) or 'possible' (p). 'Confirmed' sources were confirmed using the VLBA Calibrator Surveys (VCS; Beasley et al. 2002; Fomalont et al. 2003; Petrov et al. 2006; Kovalev et al. 2007) or the Pearson–Readhead survey (Pearson & Readhead 1988). The final catalogue contains 50 confirmed FR I objects and, 390 confirmed FR II sources.

To balance FR I and FR II sample sizes, the FRICAT catalogue of FR I radio galaxies was introduced. This catalogue, which contains 219 FR I radio sources, is a subsample of the catalogue from Best & Heckman (2012), hereafter BH12. BH12 was formed by cross-matching the optical spectroscopy catalogues produced by Brinchmann et al. (2004) and Tremonti et al. (2004) based on data release 7 of Sloan Digital Sky Survey (SDSS DR7; Abazajian et al.

2009) with the NVSS and FIRST surveys, for sources with flux densities in NVSS that were greater than 5 mJy.

To make the FRICAT sample, the authors compiled all BH12 sources with $z < 0.15$, resulting in 3357 objects. From this sample, all sources with radio emission extended more than 30 kpc from the host galaxy centre were selected, resulting in a sample of 741 sources with well-resolved morphologies. From these 741 sources, individual objects were then classified as FR I type if (1) a one-sided or two-sided jet was present (including sources with bent jets), (2) the surface brightness of the jet decreased along its length, and (3) there was little or no brightness enhancement at the end of jet.

Sources in which brightening was observed along the jet (e.g. wide-angle tailed sources; Burns, Eilek & Owen 1982), were excluded. All three FRICAT authors classified the sources independently, and an object would be added to the catalogue only if two out of three authors agreed. The final catalogue contains 219 FR I radio galaxy candidates. The hosts of the FRICAT sources were found to be luminous ($-24 \geq M_r \geq -21$) early-type galaxies with black hole masses in the range 10^8 to 3×10^9 M$_\odot$.

CoNFIG and FRICAT together form the input sample for this work. By cross-matching FR I source centroid coordinates from FRICAT and CoNFIG, we found three duplicate sources. To remain consistent with Aniyan & Thorat (2017), we did not remove this small number of duplicate objects. The final sample comprises 266 FR I sources and 390 FR II sources. We extracted the central coordinates of host in each case and used these to download FIRST images for each object using the PYTHON `astroquery` and `urllib` tools.

The `astroquery` library is an affiliated package of ASTROPY, with tools for querying astronomical web forms and data bases. We used the `astroquery.skyview.get_image_list` function, specifying central source coordinates, setting survey name to ´VLA FIRST (1.4 GHz)´ or ´NVSS´, and specifying the image scaling to be ´Linear´. With the returned list of image URLs we then used the `urllib.request.urlretrieve` function to download NVSS and FIRST images. `SkyView` by default returns each FITS image with size of 300×300 pixels (0.15×0.15 deg for FIRST images). The pixel values in these FITS images correspond to a brightness scale in units of Jy/beam.

3.2 Image pre-processing and augmentation steps

The image pre-processing in this work consists of three operations: pixel-value rescaling, image rotation, and image clipping. These are performed on all input images.

Aniyan & Thorat (2017) reported that image background noise decreased classifier performance. Having investigated various noise clipping options, they proposed a solution where ==pixel values lower than 3 times the local rms noise were set to zero,== which we followed.

After clipping, we rescaled each image following:

$$\text{Output} = \frac{\text{Input} - \text{Min}}{\text{Max} - \text{Min}} \times (255.0 - 0.0), \quad (2)$$

where `Max` and `Min` refer to the maximum and minimum pixel value in an image. `Output` and `Input` represent the final rescaled and original pixel values in the image, respectively. Rescaled images have pixel values in the range from 0 to 255. These are then saved in PNG format.

When training machine learning models, training data sets typically have sizes of order ~10 000 data samples (Aniyan & Thorat 2017). Here, the original data set contains a total of 659 source images, where ~30 per cent will serve as test samples to

Table 1. A summary of FR I, FR II images of the data set samples. The data set consists of samples for training, validation, and testing. The augmented samples are created by the process claimed in Fig. 1. In step of $1°$, FR I source images were rotated from $1°$ to $73°$. For FR II images, we rotated them from $1°$ to $50°$.

Class	Training/Validation	
	Original	Augmented
FR I	189	13 797
FR II	273	13 650
Total	462	27 447
Class	**Test**	
	Original	Augmented
FR I	80	5840
FR II	117	5850
Total	197	11 690

evaluate our trained model performance (Aniyan & Thorat 2017). Data augmentation was therefore considered to be necessary.

Considering both data set sample size and class balance, we decided to augment our data set to have around 14 000 training images and 6000 testing images for each class. Following previous practice (Aniyan & Thorat 2017), data augmentation was done by rotation of the original input source images by $1°, 2°, 3°$ etc. Table 1 provides details of our data set sample size. The final data set contains 39 796 sample images.

The final step in building our data set was to clip the image sizes. Image clipping is designed to constrain an image to its central source, yet remain large enough to identify structure. Aniyan & Thorat (2017) clipped sample images from their centres to 150×150 pixels, which is equivalent to a physical extent of 274.1 kpc at $z = 0.05$ for FIRST images. For NVSS image at the same redshift, however, image physical extent equals to 2283.8 kpc. For the CoNFIG and FRICAT samples, 96.4 per cent of them have their redshifts ≤ 0.05. We visually inspected the central source in sample images from our data set clipped in the same way, and found that the image width was suitable to recognize the sources while retaining characteristics sufficient to identify their FR morphology. Given that the sky coverage of the NVSS sample images is much larger than that of the FIRST images, the NVSS fields will inevitably include additional secondary sources. The effect of these secondary sources on training is discussed in Section 4.1.

3.3 Data formatting and division

After pre-processing, we converted our input data set images into numpy format. Target classification labels were defined as one-hot vectors: FR I samples are labelled as [1., 0.], while FR II samples have the label [0., 1.]. Section 4 discusses why such label vectors are convenient to use and will simplify the computational process when training the model. Label and image data sets were saved in two-dimensional arrays. Rather than saving images in three-dimensional arrays, we saved data input in four-dimensional arrays as they are more flexible, capable of saving both single channel images (grey-scale) and multiple channel images (e.g. RGB), and share the same format as the well-known MNIST data set (Lecun et al. 1998). We also note that data were fully shuffled before being imported into the CNN.

We split the master data set into training, validation, and test subsets. Each subset of data includes both image data and target classification labels. The training set is used to train the CNN machine learning model via back propagation. The validation set, on

the other hand, helps us to examine whether the model is overfitting the data through forward propagation. This examination takes place after every epoch training. Validation subset samples can therefore be extracted from the original training set. The test data subset is separated from the training set and validation set, with samples unseen by model before testing. This subset provides samples for evaluating the performance of the trained model when doing realistic classification. Table 1 gives an overview of the data subsets used in this work.

In the work, consistent with Aniyan & Thorat (2017), we separated training and test samples with a ratio of 70:30. We then pre-processed and applied image augmentation on all sample images. The primary training set is split into training and validation using a ratio of 80:20.

4 NETWORK ARCHITECTURE

An appropriate network architecture should consider object complexity and computational power. If necessary, transfer learning ability is a factor as well. Although simple networks may perform well for some applications (e.g. Gheller et al. 2018), classifying radio galaxy morphology requires comparatively deeper networks. Early attempts have shown that, for radio galaxy classification, simple networks perform only slightly better than random guesses (Aniyan & Thorat 2017). With fewer learnable layers, simple networks may also meet issues when using transfer learning as they have weaker expressive ability (Oquab et al. 2014). A reasonably deep network therefore becomes necessary for classifying radio galaxies and being capable of doing so using transfer learning.

Among the pre-existing architectures, the AlexNet CNN is a representative deep network (Krizhevsky et al. 2012). This is a widely used 12-layer parallel computing CNN with five convolutional layers, three max pooling layers, three fully connected layers, and a softmax readout layer. Aniyan & Thorat (2017) slightly modified this network and succeeded in classifying FR I, FR II, and Bent-tailed radio sources, with a general accuracy of above 90 per cent. Their network adopted a GPU-based implementation and was able to train 30 epochs of data in ∼70 min (Aniyan & Thorat 2017). The number of epochs in machine learning refers to the number of times that the complete training data set is imported during training.

Inspired by Aniyan & Thorat (2017), we adopt a similar but simplified 13-layer network, see Fig. 2. We initially discard the network components allowing for parallel computation. We then added another fully connected layer to reduce overfitting. Finally, instead of optimizing loss using a traditional mini-batch gradient descent optimizer (SGD; Robbins & Monro 1951) and step decay learning rate schedule, we used the adaptive mini-batch optimizer AdaGrad (Duchi, Hazan & Singer 2011) to minimize the model loss function. This optimizer algorithm works with a batch size of 100 and an initial learning rate of 0.01. The batch size refers to the number of training samples fed into the network at a time. Since we use 22 268 training samples, we import our data in 223 batches. Such a data import method is often referred to as mini-batching. The mini-batching method has typically been found to speed up the training process (Benetan & Pyzer-Knapp 2018).

Using this architecture, we observed that model validation accuracy started to saturate after 10-epoch training (Aniyan & Thorat 2017). Consequently, all training in this work is stopped after 10 epochs. The filter:node number for each layer is set to be 1:16 due to computational power limitations. The same architecture is used for both our initial model training and later transfer learning applications. Table 2 provides further network details. Notably, the
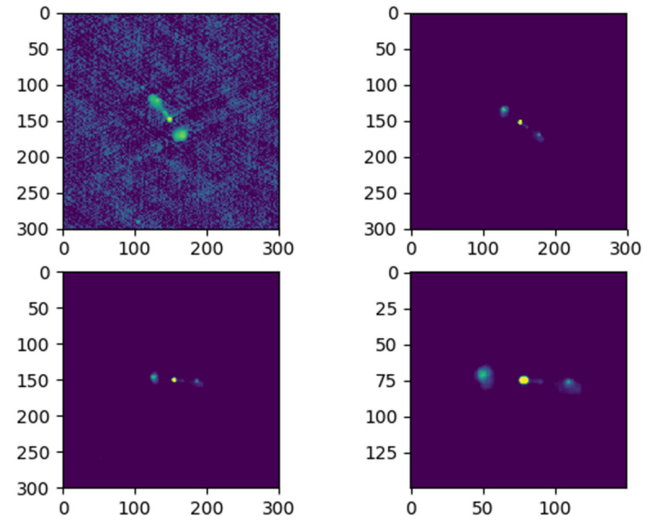

Figure 1. An example of FIRST image pre-processing and augmentation. The upper left image is the log scaled original image downloaded from SkyView. The other three images, from left to right, top to bottom are the ones experienced sigma-clipping, rotation, and centred crop. The radio source centred at the sample FIRST image is 4C 31.30, a 'confirmed' CoNFIG FR II sample. The radio galaxy host locates at (J2000) 07:45:42.13 + 31:42:52.6. The example of NVSS image pre-processing and augmentation can be seen at Figure A2.

`parameters` characteristic for each layer in Table 2 is determined by the receptive field size, the input channel number, and the depth of each layer. The `Conv1` layer, for instance, has $11 \times 11 \times 1 \times 6 + 6 = 732$ learnable parameters.

4.1 Direct classification

CNNs are capable of learning the form of the features for classification and expressing them as the values of the convolution kernel weight matrix. In the context of radio morphology classification, CNNs will learn these features from the input training samples and extract differentiable features from them. Fig. 3 shows some typical examples of training samples.

When a convolutional network trains on these samples, its initial convolutional layers will tend to learn general sample features, while lower layers are more likely to extract features specific to the data set itself. We visualize these features by plotting feature maps, which show the activation of different parts of the image (Zeiler & Fergus 2014). Fig. 4 shows feature maps for two representative samples. Features in the diagram were extracted from a randomly initialized model following 10-epoch training. These features correspond to the second and the fifth convolutional layers. Both layers implement 16 filters in total and the figure shows the first 10. Generally speaking, for the FR II source (bottom), the features learned by the second layer seemed to emphasize the existence of double edge-brightened lobes and the relative positions of the hotspots. Whereas the lower layers have learned more specific features: source outlines or the source-background relationship. This is similar to what was observed in Aniyan & Thorat (2017). All these features are saved via their model weights, and are used by the fully connected layers to make an FR binary classification.

Both NVSS and FIRST training samples were imported to train a randomly initialized model for 10 epochs each. Fig. 5 provides an overview of model learning processes. The average temporary accuracy and loss is calculated from 10 independently trained
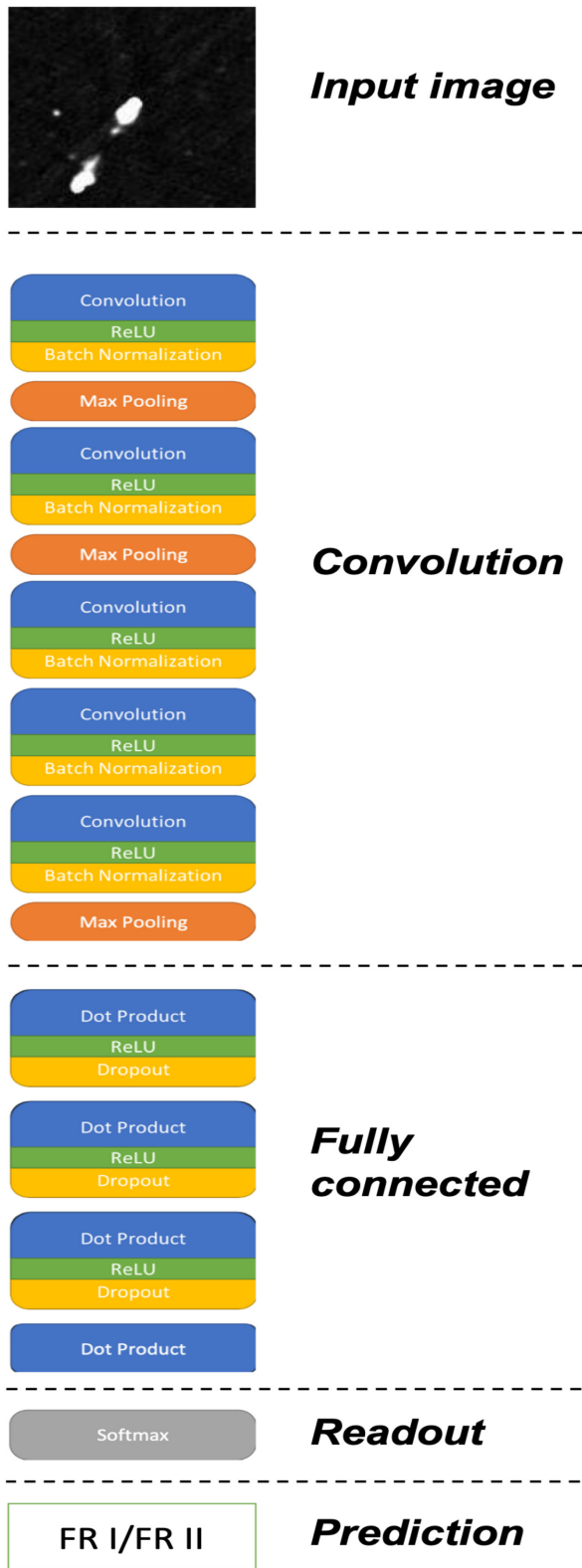
Figure 2. Network architecture adopted in the work. Blue: filters with learnable parameters; Green: activation functions; Yellow: Regularizers; Orange: Pooling layers; Grey: Softmax layer. The 13-layer architecture contains 5 convolutional layers, 3 max-pooling layers, 4 fully connected layers, and a softmax readout layer. We consider pooling and readout layers separately.

models. The standard deviation of these parameters corresponds to their error bars in the diagram. All models trained from random initializations using the Xavier uniform weight initializer ('Xavier' models hereinafter; Glorot & Bengio 2010) that experienced 10 epochs of training have a validation accuracy above 99.5 per cent and losses lower than 0.02, regardless of input sample selection.

Models trained using FIRST data as the input samples tended to have a smaller error. These models also learned more efficiently, as their training losses dropped faster. Models trained using NVSS images as input samples seemed to oscillate more frequently, which implies that these models have a higher risk of making random guesses and a lower chance of performing stable training.

The reason for this oscillation is thought to be the larger angular extent of the NVSS sample images, which contain secondary sources that might disturb the network when extracting class features. We verified this by training and testing on the same NVSS sample images but with a reduced size of 18×18 pixels, which gives an equivalent sky area to the 300×300 pixel FIRST images. Apart from the reduction in size, the network architecture remained consistent with that described previously, with the exception that no pooling layer was used. In these tests the level of oscillation was seen to decrease. We therefore conclude that the oscillation is introduced by secondary structures within the field of view, but we note that the comparatively poor resolution and sensitivity of NVSS compared to FIRST will also lead to larger uncertainties in the weights and contribute to an oscillatory effect.

4.2 Transfer learning

Though models trained from scratch perform reasonably well, whether valuable to apply transfer learning on these models had yet to be explored. There was saying that algorithms trained on data from one survey have to be trained from other surveys (Aniyan & Thorat 2017), while it is uncertain if this can be valid regardless of the choices which input survey as a starting point. Here we considered three methods:

(i) **Method 0:** Inherit the complete network architecture and weights from the pre-trained models and retrain the full network.

(ii) **Method A:** Inherit the same network and weights for all layers and retrain the fully connected layers.

(iii) **Method B:** Inherit the same network and the weights for the convolutional layers, but retrain the fully connected layers from scratch.

Here retrain means inheriting models trained with data from one survey as the network initialization, and then optimize on data from another survey. For example, we trained a network from scratch with NVSS images and then retrained the model with FIRST images afterward, and vice versa. The transfer learning methods adopted here do not change the network architecture, the early stopping criteria (10 epochs), or the training sample size and hyperparameters (e.g. batch size and primary learning rate). These methods only specify which layers to freeze, and whether to trigger the training using pre-trained weights.

The purpose of Method 0 is to examine whether transfer learning can provide a better starting point and form a better trained model. Methods A and B are intended to examine if directly inheriting features from pre-trained convolutional layers can produce comparably good results and reduce training time.

Fig. 6 shows average learning curves for each of the three methods and includes the 'Xavier' models for comparison. In general, transfer learning constrains the standard deviation compared to

Table 2. Network parameters of the classifier we adopted in the work. 'Parameters' are only available for 'Conv' and FC layers. Parameters within these layer can learn through backpropagation, while pooling and loss layers cannot learn.

| Layer	Name	Receptive field	Stride	Input channel number	Depth	Activation	Regularizer	Parameters
1	Conv 1	11×11	1	1	6	ReLU	Batch normalization	732
2	Max pooling 1	2×2	2	6				
3	Conv 2	5×5	1	6	16	ReLU	Batch normalization	2416
4	Max pooling 2	3×3	3	16				
5	Conv 3	3×3	1	16	24	ReLU	Batch normalization	3456
6	Conv 4	3×3	1	24	24	ReLU	Batch normalization	5184
7	Conv 5	3×3	1	24	16	ReLU	Batch normalization	3456
8	Max pooling 3	5×5	5					
9	Fully connected 1		1	$5 \times 5 \times 16$	256	ReLU	Dropout	102 656
10	Fully connected 2			256	256	ReLU	Dropout	65 792
11	Fully connected 3			256	256	ReLU	Dropout	65 792
12	Fully connected 4			256	2			514
13	Loss softmax							
Total parameters:								249 998
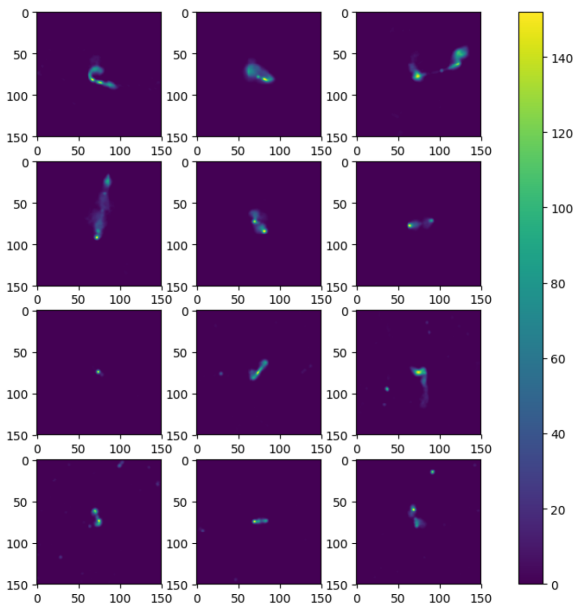
Figure 3. Examples of images used in model training. Models trained with these samples were used to classify FR morphology from test data set NVSS or FIRST images. First row: FR I samples of FIRST images; second row: FR II samples of FIRST images; third row: FR I samples of NVSS images; fourth row: FR II samples of NVSS images. The colour bar represents the linear-normalized pixel values.

direct training by at least a factor of 2. Such an effect is seen regardless of the transfer learning method applied or survey data used. In addition, transfer learning accelerated the convergence of the model in each case. Finally, the application of transfer learning provided a higher starting validation accuracy after the first epoch of training.

Among the three methods, Method 0 seems to perform best. By applying Method 0, models share highest starting points for training on both NVSS (98.7 per cent) and FIRST (99.8 per cent) after only one epoch of training. Though the accuracy gap between the 'Xavier' models and the Method 0 models gradually decreases, the models using Method 0 in fact provide better validation accuracy with final values above 99.9 per cent. In the case of independent training stability, the 'Xavier' models have a validation accuracy

standard deviation after the final epoch on order of 10^{-3}, whereas the value for models using Method 0 ranges from 10^{-5} to 10^{-4}. The time spent on training for models using Method 0 is slightly longer than training from scratch: randomly initialized models need 33 ms to train on an image, while the time required for Method 0 is about ∼34 ms. Generally speaking, each training run takes ∼124 min on our test system, see Section 1.

Comparatively, Methods A and B may require more training over a larger number of epochs as their training accuracies grow very slowly. Models using Method A or B have a validation accuracy which grows from 0.5 per cent to 0.8 per cent at each epoch, while the growth for 'Xavier' models is > 2.8 per cent. This is understandable given that the convolutional layers are frozen. However, freezing these layers reduces the time cost for model training. Both Methods A and B require only 7 ms to train each input image, ∼ 21 per cent of the time cost using Method 0. Comparing the two methods, Method A performs better that Method B due to the difference in its weight inheritance. Further analysis of classification accuracy using the three methods will be given in Section 5.1.

5 DISCUSSION

5.1 Classification accuracy

Besides validation sets, the prediction–truth comparison is another factor when evaluating a classifier. Accuracy is often considered as the primary parameter for evaluation; however, while this metric can provide an overview of model performance, this can be misleading when test samples have a significant class imbalanced.

A widely used tool for evaluating model performance on a class-wise basis is the confusion matrix (Stehman 1997), a table to visualize model performance. Fig. 7 gives an example of a confusion matrix adopted for this work. This matrix has four quadrants: True-Positive (TP), False-Negative (FN), True-Negative (TN), and False-Positive (FP). Assuming FR II morphology as 'true,' the matrix counts the FR II test samples towards TP if model prediction matches its class label identified by human.

The four quadrants can be used to further assess the predictive ability of a two class model as the basis for deriving three additional metrics: Recall (R), Precision (P), and F_1 score (Powers 2008). R is the ratio of TP and TP + FP, while P is TP /(TP + FN). In the context of classifying FR II morphology, a higher R and
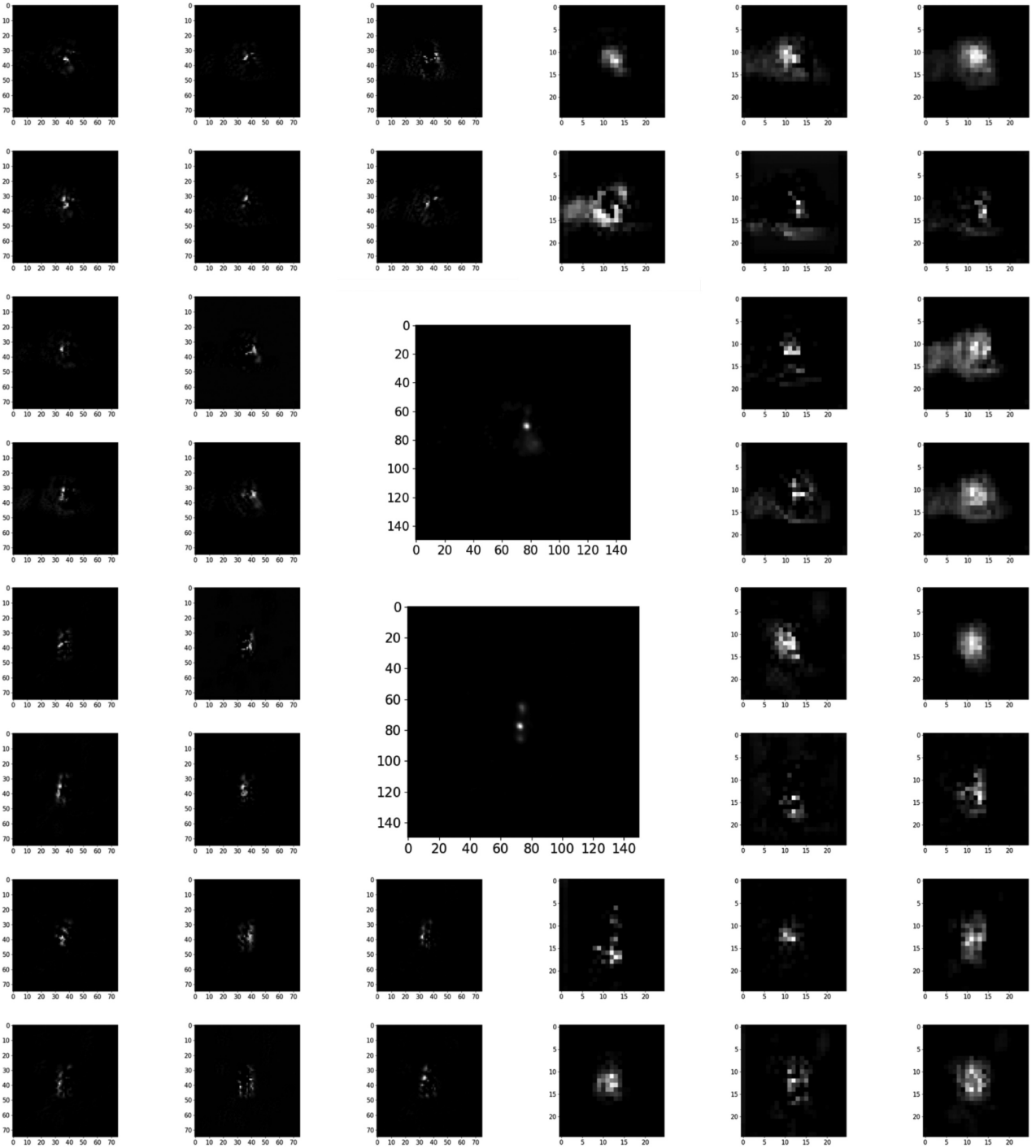
Figure 4. An example of feature maps using testing FIRST sample images. These images are produced by convolving the example source image with the first 10 filters of either the second or the fifth convolutional layer shown in Fig. 2. Upper middle: An example of FR I sources in the testing set. Lower middle: An example of FR II sources in the testing set. Upper left: Features of the example FR I source extracted by the second convolutional layer. Upper right: Features of the example FR I source extracted by the fifth convolutional layer. Lower left: Features of the example FR II source extracted by the second convolutional layer. Lower right: Features of the example FR II source extracted by the fifth convolutional layer. Source images and feature maps in the diagram are in grey-scale. Feature map examples using testing NVSS sample images can be seen at Figure A1.

P score refers to greater sensitivity and class prediction accuracy, respectively. The F_1 score,

$$F_1 = 2\frac{P \times R}{P + R}, \qquad (3)$$

can be seen as the weighted average of R and P, which provides a general assessment when identifying samples of a class. These four metrics enable us to evaluate the models trained in this work. Each model was tested with identical NVSS and FIRST image sets.
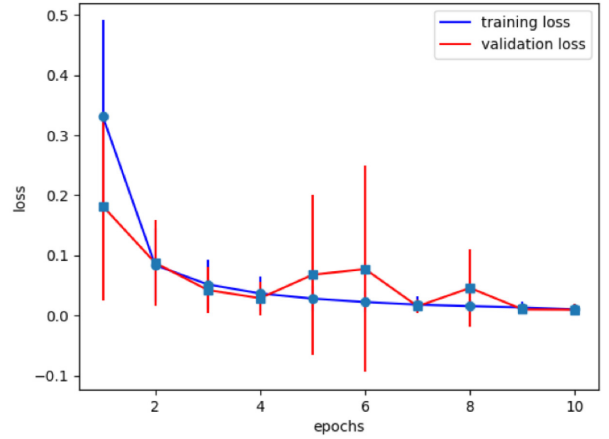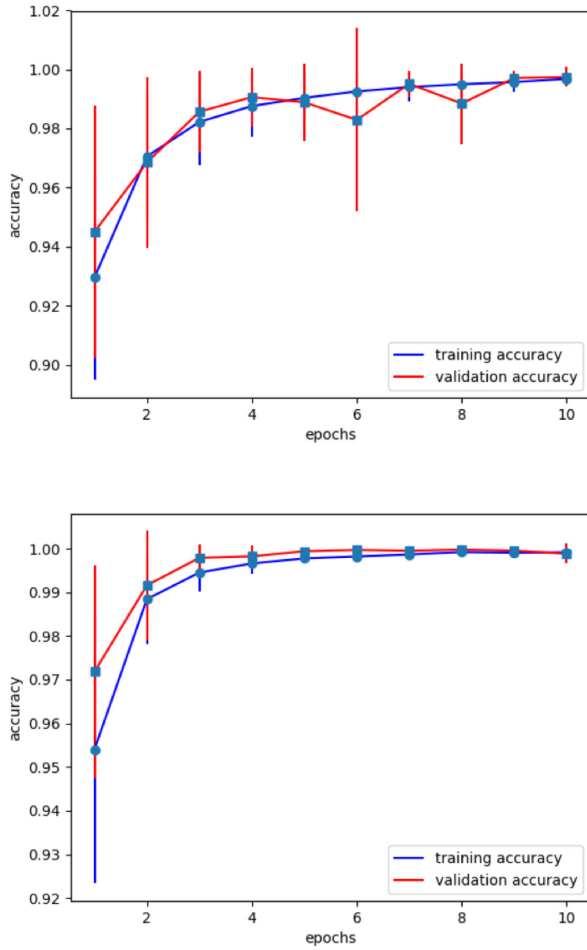
Figure 5. Upper: averaging learning and loss curve for 'Xavier' models trained on NVSS images. Lower: The same curves for models trained on FIRST images.
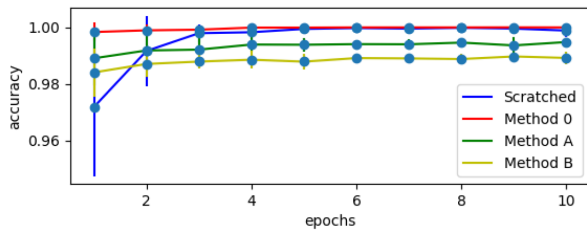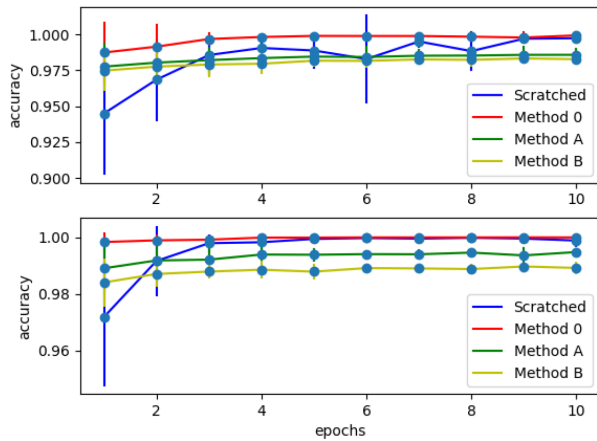

Figure 6. Upper: Model average validation accuracy curves with corresponding error bars trained with inherit NVSS images using variant methods. Blue, red, green, and yellow curves represent models trained from scratch, using Methods 0, A, and B, respectively. Lower: The same curves trained with inherited weights trained on FIRST images.
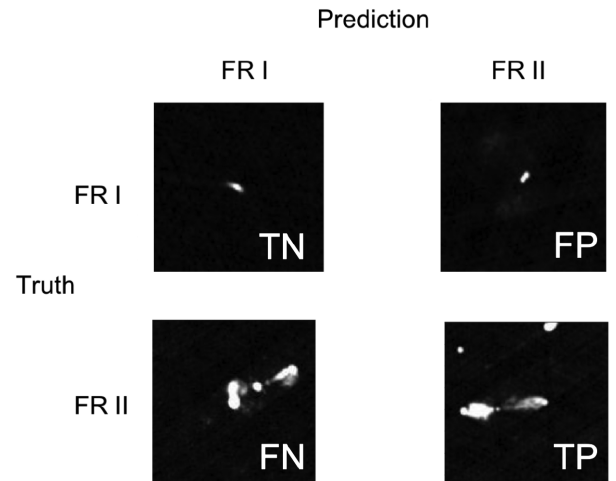

Figure 7. An example of a confusion matrix. In the context of binary classification, FR I and FR II represent false and true classes, respectively. All pre-processed FIRST images in the matrix came from the test set. Images shown in the figure share image size of 150×150 pixels.
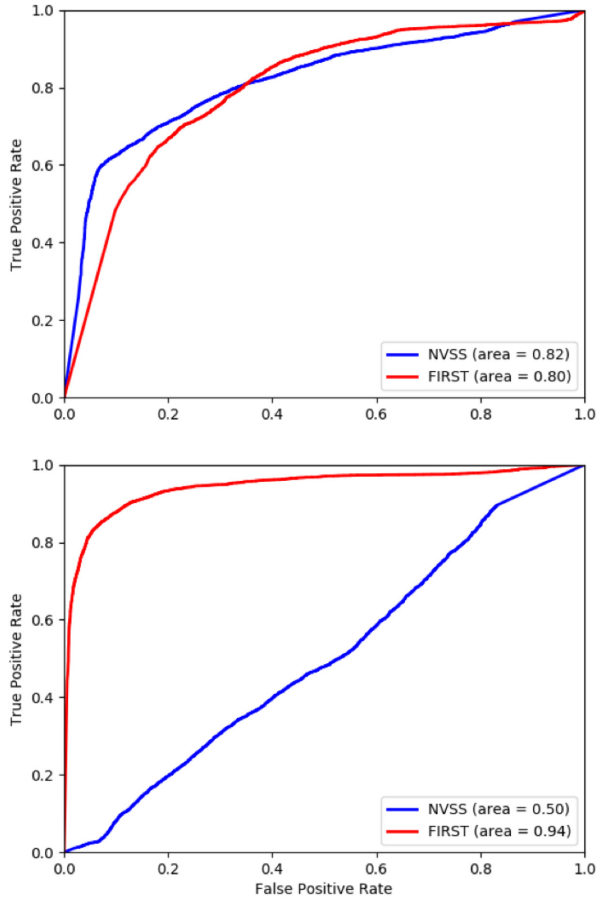
Figure 8. ROC curve for 'Xavier' models. The colours in the diagram represent the survey of the test images used to derive the curve. Blue refers to NVSS images, while red represents FIRST images. Upper: ROC curves for 'Xavier' models trained on NVSS images for 10 epochs. Lower: ROC curves for 'Xavier' models trained on FIRST images. When deriving the curves, the FR I class is assumed to be 'true', while the FR II class is considered to be 'false'.

In addition to the numerical value for each of these four metrics, the Receiver Operating Characteristic (ROC; e.g. Fig. 8) curve is also used to represent model performance. The ROC curve is mainly a useful tool when making binary classification, although multiclass variants do exist. It provides a visualization of the FP rate versus the TP rate for a number of candidate thresholds from 0 to 1. The TP rate is equal to the recall when the class being considered is the 'real' class, while equation (4) defines the FP rate:

$$\mathrm{False\,positive\,rate} = \frac{\mathrm{fp}}{\mathrm{fp} + \mathrm{tn}}. \qquad (4)$$

The ROC curve can be seen as a trade-off between the two variables, and the area under the curve (AUC) value is often used when comparing models. In such a comparison, testing on the same samples, the model with a higher AUC is considered to perform best when distinguishing classes.

5.1.1 Randomly initialized models

Depending on the choice of data sets used in training and testing, classifier performance can vary under evaluation. Fig. 8 shows the ROC curves for different models trained from scratch. Models trained with NVSS samples show similar behaviour when tested

Table 3. A summary of model performance for randomly initialized models trained for 10 epochs. Testing for models trained on one survey images adopted the test image set from the same survey.

NVSS trained	NVSS test result		FIRST test result	
	FR I	**FR II**	**FR I**	**FR II**
Recall	0.67 ± 0.01	0.87 ± 0.04	0.74 ± 0.06	0.70 ± 0.06
Precision	0.92 ± 0.02	0.54 ± 0.03	0.67 ± 0.04	0.77 ± 0.07
F1 Score	0.77 ± 0.02	0.67 ± 0.05	0.70 ± 0.08	0.73 ± 0.10
	73.0 ± 1.1		71.9 ± 2.8	
Accuracy(per cent)				
FIRST trained	**NVSS test result**		**FIRST test result**	
	FR I	**FR II**	**FR I**	**FR II**
Recall	0.49 ± 0.01	0.40 ± 0.17	0.85 ± 0.02	0.94 ± 0.04
Precision	0.92 ± 0.02	0.05 ± 0.02	0.95 ± 0.02	0.83 ± 0.04
F1 Score	0.64 ± 0.02	0.09 ± 0.06	0.90 ± 0.03	0.88 ± 0.06
	48.5 ± 1.2		89.1 ± 1.4	
Accuracy(per cent)				

on either NVSS or FIRST samples. The average AUC for testing the two sample sets is 0.80 ± 0.01 and 0.78 ± 0.02, respectively. A summary of model average AUC is shown in Table 5.

In comparison, models trained using FIRST samples show an asymmetric performance. Such models work well when classifying unlabelled (test) FIRST images, but they behave randomly when tested on NVSS images. The AUC for these models reached 0.94 ± 0.01 for the FIRST samples, while the metric for NVSS was 0.54 ± 0.05.

For other metrics the situation is similar. Table 3 summarizes the metrics described above for these models. It can be seen that models with higher AUC scores also share higher classification accuracy. Models trained using FIRST images perform best when making predictions on FIRST test images. These models generally achieve 89.1 per cent $\pm$ 1.4 per cent accuracy. When models are trained and tested on NVSS images, however, test accuracy drops to 73.0 per cent $\pm$ 1.1 per cent. Such a change might be attributed to the differences between the two surveys: sample sources in FIRST are well resolved and extended in most cases, sources in NVSS, however, are sometimes only slightly resolved and are small in the image (Fig. 3).

When models were trained on FIRST images and tested on NVSS images, however, we saw strong FR I class preference in these models. Neither model recall nor precision when classifying NVSS FR II images is higher than 0.5. Such bias also exists when models are trained with NVSS images. All randomly initialized models perform better when classifying NVSS FR I samples. Given that the training data are well balanced, with a 0.45 per cent higher number of FR I samples than FR IIs, it is unlikely that this is a consequence of class imbalance alone.

Since the model trained by Aniyan & Thorat (2017) used similar input data samples, we naively compare the results of our randomly initialized models trained on FIRST images to that work. Our precision when classifying FR I objects is ~ 95 per cent, similar to theirs. For FR II classification, our models achieved 83 per cent precision, compared to 75 per cent from their models. The average F1 score in our work is 91 per cent, 5 per cent higher than Aniyan & Thorat (2017); however, the recall of their models for classifying FR Is on the other hand is 6 per cent higher than ours. The difference between these two works may be explained in several ways.

First, the fusion model they proposed was a three-class classification model and a number of bent-tailed radio galaxies were mistakenly identified as FR II sources (Aniyan & Thorat 2017).

Table 4. A summary of averaging model accuracy. Accuracy in the table are represented in percentage. 'trained' refers to the survey data finally trained on each model. Bold implies that the method horizontally gave the best accuracy.

NVSS trained	Xavier	0	A	B
NVSS test (per cent)	73.0 ± 1.1	**73.0 ± 0.7**	69.8 ± 1.0	71.6 ± 0.8
FIRST test (%)	71.9 ± 2.8	78.4 ± 2.0	**81.1 ± 1.0**	78.3 ± 1.0
FIRST trained	**Xavier**	**0**	**A**	**B**
NVSS test (%)	48.5 ± 1.2	**50.3 ± 0.7**	46.9 ± 0.5	46.2 ± 0.6
FIRST test (%)	**89.1 ± 1.4**	87.4 ± 1.4	84.6 ± 0.6	83.8 ± 1.0

Secondly, there is a discrepancy in the number of input data samples and their distribution. When training their models, Aniyan & Thorat (2017) imported 36 000 FR Is, 32 688 FR IIs, and 25 488 Bent-tailed 'sources', whereas our complete data sample contains 39 796 samples. This might contribute to higher recall. Finally, although we imported FR Is and FR IIs from the same catalogues, the definition of FR Is between the two differs slightly: some FR Is we imported were considered as bent-tailed sources (Aniyan & Thorat 2017). This could potentially cause a difference in model performance.

5.1.2 Transfer learning models

Although a naive analysis of the transfer learning models presented in this work might initially indicate the advantages of applying these methods, it is important to consider that these models may show varying performance characteristics when applied to new or different unseen data sets. Table 4 gives a summary of the test accuracy for models trained with or without transfer learning methods, when applied to a data set different from that used for original training. Model test accuracy represents the general performance of a classifier.

The application of Method 0 boosted model classification ability when predicting NVSS images and gave the best test performance among the three methods. This boosting effect holds even if we decrease the number of training samples. We assessed this by reducing the number of training samples in the transfer learning to generate a smaller data set with < 20 per cent of the retraining samples in our primary NVSS data set. We reran Method 0 using this reduced data set and the same test data as previously. This resulted in no statistically significant change to the accuracy values for either the NVSS or FIRST test data sets.

When classifying FIRST images, however, models which inherited pre-trained weights from FIRST images and applied Method A performed best. The same result is not true in the case where the order of the survey data used for the inherited-retraining sequence was switched from FIRST to NVSS.

The AUC values also provide further detail. Models which used Method 0 showed comparatively stronger expressive ability than randomly initialized ones. Adopting Methods A and B produced a similar effect when inheriting weights from models pre-trained on FIRST images. Such a phenomenon, however, lost its efficacy if inheriting weights from models initially trained on NVSS images. This can perhaps be explained by the difference between the two surveys. Images of many sources seen in the FIRST survey possess richer structural information than provided by NVSS. This is an important factor when considering the application of pre-trained
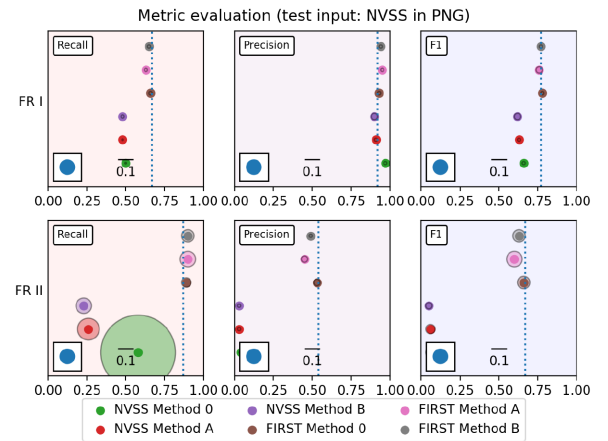

Figure 9. A summary of metric evaluation for models applied transfer learning and tested on NVSS images. 'NVSS' or 'FIRST' shown in the legend box implies that, when applying transfer learning, the pre-trained model weights were trained on the named survey. In the diagram, radius of the circles accounts for the standard deviations of their respective metrics. Dashed vertical lines refer to average metrics for the Xavier models trained and tested on NVSS images.
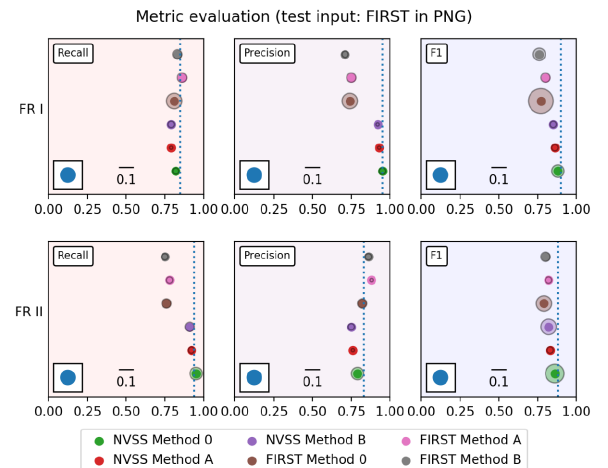

Figure 10. A summary of metric evaluation for models applied transfer learning and tested on FIRST images. Models evaluated in the diagram are the same as Fig. 9. The meanings of symbols and texts in the diagram are consistent to Fig. 9 as well. Dashed vertical lines refer to average metrics for the Xavier models trained and tested on FIRST images.

models from existing surveys to new data from next-generation telescopes such as ASKAP, MeerKAT, and the SKA.

In addition to accuracy and AUC values, we also evaluated these transfer learning models further by measuring their recall, precision, and F1 score for each class. Figs 9 and 10 show how transfer learning models behave when classifying either FR Is or FR IIs on NVSS and FIRST images. We note that models consistently identified most test samples as FR Is if they were retrained with FIRST inputs and tested using an NVSS test data set. When models are retrained with NVSS images we found that introducing transfer learning improved both FR I and FR II classification for NVSS images. The identification of FR II objects typically reached 88 per cent precision using Method A. This is ∼ 5 per cent higher than models trained with FIRST images directly. In the context of FR I classification using FIRST samples, the highest achieved precision value was 95 per cent. Such

Table 5. A summary of averaging model AUC. 'trained' refers to the survey data finally trained on each model. Bold implies that the method horizontally gave the highest AUC.

NVSS trained	Xavier	0	A	B
NVSS test	0.80 ± 0.01	$\mathbf{0.81 \pm 0.01}$	0.80 ± 0.01	0.81 ± 0.00
FIRST test	0.78 ± 0.02	0.86 ± 0.01	$\mathbf{0.88 \pm 0.01}$	0.83 ± 0.01
FIRST trained	**Xavier**	**0**	**A**	**B**
NVSS test	0.54 ± 0.05	$\mathbf{0.59 \pm 0.02}$	0.57 ± 0.01	0.53 ± 0.01
FIRST test	0.94 ± 0.01	$\mathbf{0.94 \pm 0.00}$	0.93 ± 0.00	0.92 ± 0.00

Table 6. A summary of averaging model accuracy. Model inputs considered in the diagram are in JPEG image format. Accuracy in the table are represented inpercentage. 'trained' refers to the survey data finally trained on each model. Bold implies that the method horizontally gave the best accuracy.

NVSS trained	Xavier	0	A	B
NVSS test(%)	82.1 ± 4.4	$\mathbf{83.9 \pm 1.2}$	81.8 ± 0.6	82.6 ± 0.6
FIRST test(%)	65.5 ± 7.4	$\mathbf{73.7 \pm 2.0}$	65.9 ± 1.0	66.0 ± 1.0
FIRST trained	**Xavier**	**0**	**A**	**B**
NVSS test(%)	$\mathbf{57.8 \pm 1.9}$	57.7 ± 1.2	56.6 ± 0.5	57.2 ± 1.7
FIRST test(%)	$\mathbf{90.1 \pm 1.0}$	89.7 ± 0.8	87.6 ± 0.7	87.2 ± 0.9

precision could either be achieved through direct training or by using Method 0 when training with FIRST images.

These results imply that choice of method is a trade-off. Applying Method 0 would strengthen model stability and boost the performance of a model if one wished to apply the model to both NVSS and FIRST images. Method A can boost prediction precision when identifying FR IIs on FIRST images. If not training FIRST images from scratch, applying Method A can make the most precise prediction on FIRST images. Method B can be seen as an alternative option in the case where computational power is constrained and one wants a quickly trained model which makes a reasonably good prediction.

5.1.3 Influence of input image format

The input images used for model training and testing described in Section 5.1.1 and Section 5.1.2 are processed in PNG image format. When saving images, such a format converts the value of each pixel to an integer in a lossless fashion.

In addition to PNG format, many classifiers also accept images in JPEG format. The advantage of using JPEG images is that they can require smaller storage volume and have enhanced smoothness. However, when converting image arrays to JPEG format, the images are compressed and there is information loss.

The influence of input image format on model performance has not been addressed in the context of radio galaxy classification. However, the issue of archival data storage for the next generation of radio telescopes may have implications for training data availability. In order to investigate the effect of image format we repeated the image pre-processing, data augmentation, model training, and transfer learning processes described above using images input in JPEG format in order to compare the resulting model outcomes with those using PNG inputs.
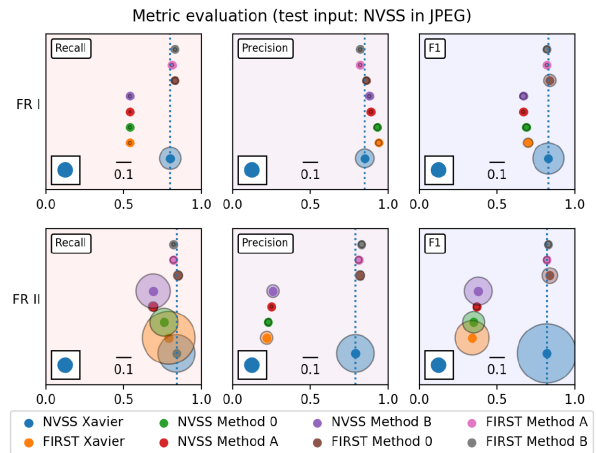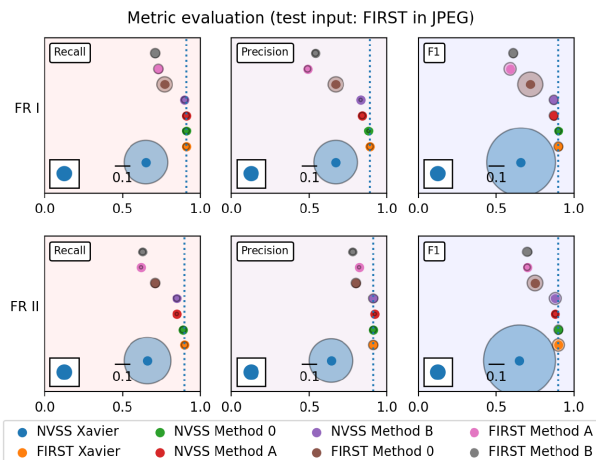
Figure 11. A summary of metric evaluation for models applied transfer learning and tested on NVSS images in JPEG input format. For transfer learning models, 'NVSS' or 'FIRST' shown in the legend box implies that, the pre-trained model weights were trained on the named survey. For 'Xavier' models, however, survey name refer to the survey data used in model training. In the diagram, the radius of the circles accounts for the standard deviations of their respective metrics. Dashed vertical lines, on the other hand, represent the average metrics for Xavier models trained and tested on NVSS images.

Figure 12. A summary of metric evaluation for models applied transfer learning and tested on FIRST images in JPEG input format. Models evaluated in the diagram are the same as Fig. 11. The meanings of symbols and texts in the diagram are consistent to Fig. 11. Dashed vertical lines, on the other hand, represent the average metrics for 'Xavier' models trained and tested on FIRST images.

Table 6 summarizes model performances using image inputs in JPEG format. Models using JPEG inputs show stronger identification ability. Comparing with Table 4, classifiers primarily trained with NVSS images showed a 9 per cent accuracy improvement when classifying NVSS test sets. For those models trained with FIRST images, however, classification accuracy is boosted for both NVSS and FIRST test data sets.

When considering Figs 11 and 12, it can be seen that the F1 score of Xavier models when classifying NVSS FR Is and FR IIs shares a common improvement. Typically, these models have their FR II identification ability strengthened significantly. The precision of FR II classification on NVSS test images reached 79 per cent, while the
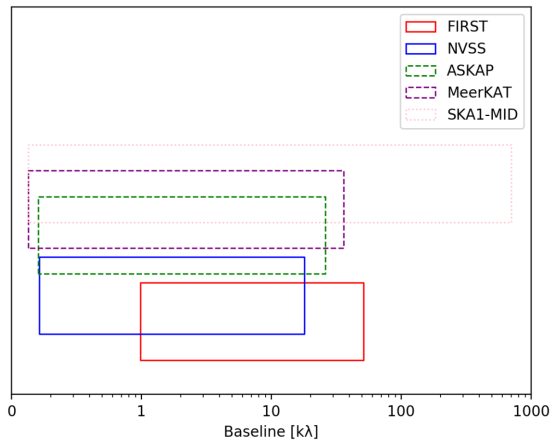
Figure 13. A summary of spatial scales for several radio telescopes/surveys in units of kilo-lambda (kλ). Solid: finished radio surveys. Dashed: radio telescopes (almost) finish construction. Dotted: telescope would be built in the future. Spatial scales shown in the diagram are converted from telescope baselines in units of km. The frequency adopted when doing the conversion is 1.4 GHz for FIRST, NVSS, MeerKAT, and SKA1-MID. We adopted 1.3 GHz for ASKAP specifically for its EMU survey (Norris et al. 2011). FIRST was observed using the VLA B-configuration of the VLA (Becker et al. 1995), while NVSS adopted the more compact D and DnC configurations of the same array (Condon et al. 1998). ASKAP have minimum and maximum baseline of 37 m and 6 km, respectively (Johnston et al. 2008; Serra et al. 2015). Baselines of MeerKAT ranges from 29 m to 7 km (Jonas & MeerKAT Team 2016). Finally, SKA1-MID is expected to have 150 km maximum baseline. The shortest baseline of SKA1-MID here is the same as MeerKAT, as MeerKAT will finally become a part of SKA1-MID core (Braun et al. 2015).

number when using PNG input was only 54 per cent. Nevertheless, when considering FIRST images, the models showed a balanced but relatively smaller recall, precision, and F1 score.

The `Xavier` models trained with FIRST images also showed general improvement when identifying NVSS images. Though the issue of FR I preference still exists, recall of FR IIs classification increased by 39 per cent compared to that using PNG inputs. This implies that by using JPEG images, the classifier achieved a higher sensitivity for identifying FR IIs.

The JPEG-based results also echo the transfer learning outcomes seen using PNG format. No matter which method was applied, models inheriting weights trained on FIRST images and then retrained on NVSS images make a more accurate prediction. Typically, by applying Method 0, models work optimally for classifying both NVSS and FIRST images. When transfer learning models inherited weights from models trained on FIRST images, their performance is similar to that using PNG inputs.

In spite of the similarities, we note that there are two other phenomena worth mentioning. The first is that the difference between randomly initialized models and those using Method 0 are reduced when using JPEG inputs. The accuracy difference between the two is less than 0.5 per cent, while the difference is larger than 1.5 per cent using PNG formatted inputs. The second phenomenon is caused by applying Method A. The application of Method A no longer makes the best FIRST prediction if retrained on NVSS images in JPEG format. If one adopted image input in JPEG when learning and testing, Methods 0 and B would become better options.

Why changing image format leads to overall model performance enhancement is not immediately obvious. To explain it, we consider the different input images from the perspective of their information

Table 7. A summary of Shannon entropy measurement for image inputs in different formats. Shannon entropy for inputs in FITS, JPEG, and PNG format have all experienced image pre-processing.

NVSS inputs	FITS	PNG	JPEG
CoNFIG FR I	0.32 ± 0.12	0.18 ± 0.11	0.22 ± 0.14
CoNFIG FR II	0.28 ± 0.03	0.15 ± 0.7	0.19 ± 0.1
FRICAT	0.28 ± 0.03	0.20 ± 0.06	0.28 ± 0.11
FIRST inputs	**FITS**	**PNG**	**JPEG**
CoNFIG FR I	0.25 ± 0.16	0.15 ± 0.09	0.19 ± 0.11
CoNFIG FR II	0.14 ± 0.09	0.06 ± 0.05	0.07 ± 0.07
FRICAT	0.09 ± 0.05	0.07 ± 0.03	0.19 ± 0.09

content. We do this by evaluating the Shannon entropy of input images in FITS, PNG, and JPEG formats. It is noted that all images have experienced sigma-clipping.

Shannon entropy refers to the averaged self-information content of a data set (Shannon & Weaver 1949). Self-information can be defined as the probability that a stochastic source of noise has produced the information in the data set. Equation (5) gives the mathematical definition of Shannon entropy, S,

$$S = -\sum p_k \log p_k \qquad (5)$$

where p_k represents the normalized pixel values considered as probabilities. For this work, we adopt 2 as the logarithmic base when measuring Shannon entropy.

By definition, inputs with lower Shannon entropy have smaller variation. Also, since the image inputs in this work are normalized to the same pixel range (0–255), an image with high Shannon entropy should have a weakly concentrated pixel value distribution. In other words, a model would find it easier to learn image pixel value gradients if the same image had higher Shannon entropy.

We compared mean Shannon entropy between inputs in different data formats from different surveys. Table 7 provides a summary of these entropy measurements. In general, NVSS sample images have higher Shannon entropy than FIRST sample images. When we take Table 4 and Table 6 into account, we find that most models retrained with FIRST images tended to have a smaller standard deviation in accuracy.

When converting input images from the CoNFIG catalogue from FITS format to JPEG or PNG, we found that their Shannon entropy consistently dropped. Images in PNG format have the lowest mean Shannon entropy of all three formats. In the context of classification, models seem to make a more accurate predictions if the input data shares higher mean Shannon entropy. Since inputs in FITS format have the highest Shannon entropy, it is recommended that future networks use FITS inputs for both training and testing machine learning models.

Regardless of survey or catalogue, FR II inputs experiencing format conversion show higher fractional loss of entropy than FR Is. However, we did not see a relationship between this loss and model performance. When we train and test model using data from the same survey, Recall, Precision, and F1 score differences between the two classes are less than 10 per cent. Such differences become more apparent when testing on a different survey, but the self-information imbalance between the two classes is not sufficient to explain the difference. When applying transfer learning, models continued to give comparative or more accurate FR II and FR I predictions on FIRST and NVSS images, respectively.

Overall, how image formats affects model performance still requires further investigation. Whether Shannon entropy can be seen as an evaluation factor also needs examination in the future.

6 THE APPLICATION OF TRANSFER LEARNING TO FUTURE RADIO SURVEYS

Traditionally, radio galaxy classification has been done by visual inspection, sometimes facilitated by measurement of host-hotspot relative positions. Such a method was practical due to the modest sample size of archival catalogues. However, soon next-generation radio catalogues such as that from the EMU survey (Norris et al. 2011), will discover millions of radio sources waiting for visual inspection.

In order to overcome the difficulties of classifying these sources by eye, recent studies have focused on developing machine learning based automated methods to classify radio source morphologies based on specific radio surveys. In this paper, we have introduced the next step in the use of these methods and explored the possibility to boost model performance by applying transfer learning.

Our approaches achieved over 90.1 per cent and 83.9 per cent in terms of classification accuracy when testing on FIRST and NVSS images, comparable with other recent state-of-the-art results. Depending on the transfer learning method used, we have demonstrated that transfer learning models can result in even higher model accuracies or save training time by up to 79 per cent.

A key result from this work is that inheriting model weights pre-trained on higher resolution survey data, e.g. FIRST, can boost model performance when retraining with lower resolution survey images, e.g. NVSS. However, we found that the reverse situation, whereby weights inherited from models trained on lower resolution data are retrained on higher resolution data, is detrimental to model performance. This is of particular relevance for future radio surveys, where machine learning weights inherited from models trained on archival data may be used to initiate classifiers for previously unseen data. Fig. 13 summarizes the baseline ranges of the NVSS and FIRST surveys, along with the ranges of ASKAP, MeerKAT, and SKA1-MID. These three telescopes are capable of making observations at 1.3–1.4 GHz, similar to FIRST and NVSS which were made at 1.4 GHz. It can be seen that there are considerable spatial scale overlap between MeerKAT, ASKAP, and the surveys considered in this work.

The higher resolution of the FIRST survey, relative to both MeerKAT and ASKAP as well as NVSS, suggests the potential for successful transfer learning approaches to machine learning classification of the survey data from these next-generation telescopes; however, the significantly improved resolution of SKA1-MID in comparison suggests that further investigations must be made before the advantages of transfer learning can be used there.

ACKNOWLEDGEMENTS

REFERENCES

Abazajian K. N. et al., 2009, ApJS, 182, 543

Ackermann S., Schawinski K., Zhang C., Weigel A. K., Turp M. D., 2018, MNRAS, 479, 415
Alger M. J. et al., 2018, MNRAS, 478, 5547
Aniyan A. K., Thorat K., 2017, ApJS, 230, 20
Banfield J. K. et al., 2015, MNRAS, 453, 2326
Beasley A. J., Gordon D., Peck A. B., Petrov L., MacMillan D. S., Fomalont E. B., Ma C., 2002, ApJS, 141, 13
Becker R. H., White R. L., Helfand D. J., 1995, ApJ, 450, 559
Benetan M., Pyzer-Knapp E. O., 2018, Practical Considerations for Probabilistic Backpropagation, Third workshop on Bayesian Deep Learning (NeurIPS 2018). Montréal, Canada
Best P. N., Heckman T. M., 2012, MNRAS, 421, 1569 (BH12)
Bock D. C.-J., Large M. I., Sadler E. M., 1999, AJ, 117, 1578
Braun R. et al., 2015, SKA1 LEVEL 0 SCIENCE REQUIREMENTS. SKA-TEL-SKO-0000007-Rev02
Bridle A. H., Perley R. A., 1984, ARA&A, 22, 319
Brinchmann J., Charlot S., White S. D. M., Tremonti C., Kauffmann G., Heckman T., Brinkmann J., 2004, MNRAS, 351, 1151
Burns J. O., Eilek J. A., Owen F. N., 1982, in Heeschen D. S., Wade C. M., eds, IAU Symposium 97, vol. 97. Springer, Netherlands, p. 45
Butenko A. V., Tyul'bashev S. A., 2016, Astron. Rep., 60, 718
Capetti A., Massaro F., Baldi R. D., 2017, A&A, 598, A49
Carilli C. L., Rawlings S., 2004, New Astron. Rev., 48, 979
Changpinyo S., Sandler M., Zhmoginov A., 2017, preprint (arXiv:1702.06257)
Condon J. J., Cotton W. D., Greisen E. W., Yin Q. F., Perley R. A., Taylor G. B., Broderick J. J., 1998, AJ, 115, 1693
Connor L., van Leeuwen J., 2018, AJ, 156, 256
Cotter G., Rawlings S., Saunders R., 1996, MNRAS, 281, 1081
Dabhade P., Gaikwad M., Bagchi J., Pandey-Pommier M., Sankhyayan S., Raychaudhury S., 2017, MNRAS, 469, 2886
Domínguez Sánchez H. et al., 2019, MNRAS, 484, 93
Duchi J., Hazan E., Singer Y., 2011, J. Mach. Learn. Res., 12, 2121
Duda R. O., Hart P. E., Stork D. G., 2012, Pattern Classification. John Wiley & Sons, New York, NY, USA
Fanaroff B. L., Riley J. M., 1974, MNRAS, 167, 31P
Fomalont E. B., Petrov L., MacMillan D. S., Gordon D., Ma C., 2003, AJ, 126, 2562
Gendre M. A., Wall J. V., 2008, MNRAS, 390, 819
Gendre M. A., Best P. N., Wall J. V., 2010, MNRAS, 404, 1719
Gheller C., Vazza F., Bonafede A., 2018, MNRAS, 480, 3749
Glorot X., Bengio Y., 2010, in Yee W. T., Mike T., eds, Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, PMLR, 9, p. 249
Glorot X., Bordes A., Bengio X., 2011, in Geoffrey G., David D., Miroslav D., eds, Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Vol. 15. p. 315
Goodfellow I., Bengio Y., Courville A., 2016, Deep Learning. MIT Press, Cambridge, MA, USA, p. 196
Grefenstette E., Blunsom P., Freitas N., Hermann K. M., 2014, preprint (arXiv:1404.7296)
Hahnloser R. H. R. et al., 2000, Nature, 405, 947
Howard A. G., Zhu M., Chen B., Kalenichenko D., Wang W., Weyand T., Andreetto M., Adam H., 2017, preprint (arXiv:1704.04861)
Huang J., Zhou W., Zhang Q., Li H., Li W., 2018, preprint (arXiv:1801.10111)
Héas P., 2018, preprint (arXiv:1806.01916)
Ioffe S., Szegedy C., 2015, preprint (arXiv:1502.03167)
Ishwara-Chandra C. H., Saikia D. J., 1999, MNRAS, 309, 100
Johnston S. et al., 2008, Exp. Astron., 22, 151
Jonas J., MeerKAT Team 2016, in Russ T., Fernando C., Lerothodi L., Kavilan M., eds, Proceedings of MeerKAT Science: On the Pathway to the SKA, Sissa Medialab srl Partita, Italy IVA:01097780322, Stellenbosch, South Africa (MeerKAT2016)
Kapinska A. D., Hardcastle M., Jackson C. et al., 2015, Advancing Astrophysics with the Square Kilometre Array (AASKA14). Sissa Medialab srl Partita, Italy IVA:01097780322, p. 173

Kimura A., Takahashi I., Tanaka M., Yasuda N., Ueda N., Yoshida N., 2017, preprint (arXiv:1711.11526)

Kohonen T., 1982, Biol. Cybern., 43, 59

Kovalev Y. Y., Petrov L., Fomalont E. B., Gordon D., 2007, AJ, 133, 1236

Krizhevsky A., Sutskever I., Hinton G. E., 2012, Advances in Neural Information Processing Systems. MIT Press, Cambridge, MA, p. 1097

Kuźmicz A., Jamrozy M., 2012, MNRAS, 426, 851

Lara L., Cotton W. D., Feretti L., Giovannini G., Marcaide J. M., Márquez I., Venturi T., 2001, A&A, 370, 409

Leahy J. P., 1993, in Röser H.-J., Meisenheimer K., eds, Lecture Notes in Physics, Vol. 421, Jets in Extragalactic Radio Sources. Springer-Verlag, Berlin, p. 1

Leahy J. P., Bridle A. H., Strom R. G., 1996, in Ekers R. D., Fanti C., Padrielli L., eds, Proc. IAU Symp., Vol. 175, Extragalactic radio sources. Kluwer, Dordrecht, p. 157

Lecun Y., Bottou L., Bengio Y., Haffner P., 1998, Proc. IEEE, 86, 2278

Li X., Chen S., Hu X., Yang J., 2018, preprint (arXiv:1801.05134)

Lukic V., Brüggen M., Banfield J. K., Wong O. I., Rudnick L., Norris R. P., Simmons B., 2018, MNRAS, 476, 246

Ma Z. et al., 2019, ApJS, 240, 34

Machalski J., Jamrozy M., 2006, A&A, 454, 95

Machalski J., Jamrozy M., Zola S., 2001, A&A, 371, 445

Machalski J., Jamrozy M., Zola S., Koziel D., 2006, A&A, 454, 85

Machalski J., Koziel-Wierzbowska D., Jamrozy M., 2007, Acta Astron., 57, 227

Maitra D. S., Bhattacharya U., Parui S. K., 2015, 13th International Conference on Document Analysis and Recognition (ICDAR). IEEE Computer Society DC, Washington, USA, p. 1021

Matsugu M., Mori K., Mitari Y., Kaneda Y., 2003, Neural Netw., 16, 555

Mauch T., Murphy T., Buttery H. J., Curran J., Hunstead R. W., Piestrzynski B., Robertson J. G., Sadler E. M., 2003, MNRAS, 342, 1117

McAlpine K. et al., 2015, Advancing Astrophysics with the Square Kilometre Array (AASKA14), Sissa Medialab srl Partita, Italy IVA: 01097780322, p. 83

Metchalf R. B. et al., 2018, A&A, 625, A119

Minsky M. L., Papert S. A., 1988, Perceptrons: Expanded Edition, MIT Press, Cambridge, MA, USA

Norris R. P. et al., 2011, Publ. Astron. Soc. Aust., 28, 215

Oquab M., Bottou L., Laptev I., Sivic J., 2014, in Proc. IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Washington, DC, USA, p. 1717

Pearson T. J., Readhead A. C. S., 1988, ApJ, 328, 114

Perucho M., Martí J. M., Laing R. A., Hardee P. E., 2014, MNRAS, 441, 1488

Petrov L., Kovalev Y. Y., Fomalont E. B., Gordon D., 2006, AJ, 131, 1872

Planck Collaboration I, 2018, preprint (arXiv:1807.06205)

Polsterer K. L., Gieseke F., Igel C., 2015, in Taylor A. R., Rosolowsky E., eds, ASP Conf. Ser., Vol. 495, Astronomical Data Analysis Software an Systems XXIV (ADASS XXIV). Astron. Soc. Pac., San Francisco, p. 81

Powers D., 2008, Mach. Learn. Technol., 2, 37

Pratt L. Y., 1993, in Stephen J. H., Jack D. C., Lee C. G., eds, Advances in Neural Information Processing Systems 5. M organ Kaufmann Publishers Inc., San Francisco, CA, USA, p. 204

Robbins H., Monro S., 1951, Ann. Math. Stat., 22, 400

Rosenblatt F., 1961, Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms, Spartan Books, Washington DC

Russakovsky O. et al., 2015, International Journal of Computer Vision, 115, 211

Saripalli L., Hunstead R. W., Subrahmanyan R., Boyce E., 2005, AJ, 130, 896

Schoenmakers A. P., de Bruyn A. G., Röttgering H. J. A., van der Laan H., 2001, A&A, 374, 861

Sebastian R., 2016, preprint (arXiv:1609.04747)

Serra P. et al., 2015, MNRAS, 452, 2680

Shannon C. E., Weaver W., 1949, Communication Theory of Secrecy Systems. University of Illinois Press, Urbana

Skrutskie M. F. et al., 2006, AJ, 131, 1163

Solovyov D. I., Verkhodanov O. V., 2011, Astrophys. Bull., 66, 416

Song Y., Schwing A. G., Zemel R. S., Urtasun R., 2016, preprint (arXiv:1511.06411)

Sonntag D., Barz M., Zacharias J., Stauden S., Rahmani V., Fóthi Á., Lőrincz A., 2017, preprint (arXiv:1709.01476)

Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov O., 2014, J. Mach. Learn. Res., 15, 1929

Stehman S. V., 1997, Remote Sens. Environ., 62, 77

Subrahmanyan R., Saripalli L., Hunstead R. W., 1996, MNRAS, 279, 257

Szegedy et al., 2015, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Institute of Eletrical and Electronics Engineers (IEEE), p. 1

Tremonti C. A. et al., 2004, ApJ, 613, 898

White R. L., Becker R. H., Helfand D. J., Gregg M. D., 1997, ApJ, 475, 479

Wu C. et al., 2019, MNRAS, 482, 1211

Yosinski J., Clune J., Bengio Y., Lipson H., 2014, the 27th International Conference on Neural Information Processing Systems, Vol. 2. MIT Press, Cambridge, MA, USA, p. 3320

Zeiler M. D., Fergus R., 2014, Computer Vision - ECCV2014, vol. 8689. Springer, Cham

APPENDIX A: SUPPLEMENTARY IMAGES

Figure A1. An example of feature maps using testing NVSS sample images. These images are produced by convolving the example source image with the first 10 filters of either the second or the fifth convolutional layer shown in Fig. 2. Upper middle: An example of FR I sources in the testing set. Lower middle: An example of FR II sources in the testing set. Upper left: Features of the example FR I source extracted by the second convolutional layer. Upper right: Features of the example FR I source extracted by the fifth convolutional layer. Lower left: Features of the example FR II source extracted by the second convolutional layer. Lower right: Features of the example FR II source extracted by the fifth convolutional layer. Source images in the diagram are in grey-scale.
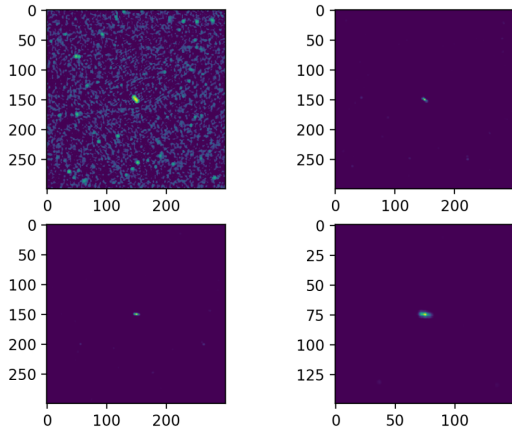
Figure A2. An example of NVSS image pre-processing and augmentation. The upper left image is the log scaled original image downloaded from SkyView. The other three images, from left to right, top to bottom are the ones experienced sigma-clipping, rotation, and centred crop. The radio source centred at the sample NVSS image is 4C 31.30, a 'confirmed' CoNFIG FR II sample. The radio galaxy host locates at (J2000) 07:45:42.13 + 31:42:52.6.

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