Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **514,** 2599–2613 (2022) https://doi.org/10.1093/mnras/stac1135
Advance Access publication 2022 April 29

Radio Galaxy Zoo: using semi-supervised learning to leverage large unlabelled data sets for radio galaxy classification under data set shift

Inigo V. Slijepcevic [1]★ Anna M. M. Scaife [1,2] Mike Walmsley [1] Micah Bowles,[1] O. Ivy Wong [3,4,5] Stanislav S. Shabala [5,6] and Hongming Tang [7]

[1]*Jodrell Bank Centre for Astrophysics, Department of Physics and Astronomy, University of Manchester, Manchester M13 9PL, UK*
[2]*The Alan Turing Institute, Euston Road, London NW1 2DB, UK*
[3]*CSIRO Space and Astronomy, PO Box 1130, Bentley, WA 6102, Australia*
[4]*ICRAR-M468, University of Western Australia, Crawley, WA 6009, Australia*
[5]*ARC Centre of Excellence for All Sky Astrophysics in 3 Dimensions (ASTRO 3D), Australian National Univeristy, Stromlo, ACT 2611, Australia*
[6]*School of Natural Sciences, University of Tasmania, Private Bag 37, Hobart, Tas 7001, Australia*
[7]*Department of Astronomy, Tsinghua University, Beijing 100084, China*

Accepted 2022 April 20. Received 2022 April 7; in original form 2022 January 28

ABSTRACT

In this work, we examine the classification accuracy and robustness of a state-of-the-art semi-supervised learning (SSL) algorithm applied to the morphological classification of radio galaxies. We test if SSL with fewer labels can achieve test accuracies comparable to the supervised state of the art and whether this holds when incorporating previously unseen data. We find that for the radio galaxy classification problem considered, SSL provides additional regularization and outperforms the baseline test accuracy. However, in contrast to model performance metrics reported on computer science benchmarking data sets, we find that improvement is limited to a narrow range of label volumes, with performance falling off rapidly at low label volumes. Additionally, we show that SSL does not improve model calibration, regardless of whether classification is improved. Moreover, we find that when different underlying catalogues drawn from the same radio survey are used to provide the labelled and unlabelled data sets required for SSL, a significant drop in classification performance is observed, highlighting the difficulty of applying SSL techniques under data set shift. We show that a class-imbalanced unlabelled data pool negatively affects performance through prior probability shift, which we suggest may explain this performance drop, and that using the Fréchet distance between labelled and unlabelled data sets as a measure of data set shift can provide a prediction of model performance, but that for typical radio galaxy data sets with labelled sample volumes of $\mathcal{O}(10^3)$, the sample variance associated with this technique is high and the technique is in general not sufficiently robust to replace a train–test cycle.

Key words: methods: data analysis – methods: statistical – radio continuum: galaxies.

1 INTRODUCTION

Radio galaxies are a subset of active galactic nuclei (AGNs) that typically exhibit a pair of roughly symmetric jets that are usually pointed in opposite directions, although high ram pressure can cause some examples to exhibit a 'bent-tailed' structure (Mguda et al. 2015). The jets are powerful emitters across the electromagnetic spectrum and their radio emission is dominated by synchrotron emission from ultrarelativistic electrons (Hardcastle & Croston 2020). Historically, radio galaxy morphologies have been assigned to various categories. The most persistent amongst these is the Fanaroff–Riley classification (Fanaroff & Riley 1974), with radio galaxies split into the Fanaroff–Riley type I (FRI) and Fanaroff–Riley type II (FRII) categories based on the distance between the brightest point on each lobe as a proportion of the total length of the source (Fanaroff & Riley 1974). FRI (FRII) sources have lower (higher) brightness lobes, whose brightness typically decrease (increase) as distance to the centre increases.

Although progress has been made in relating the two FR classes to the dynamics and energetics of the sources (e.g. Saripalli 2012; Turner & Shabala 2015; Ineson et al. 2017; Hardcastle 2018), our understanding of the causal relationship between a source's FR classification and physical environment/properties is incomplete with a number of outstanding questions that still remain. In particular many existing studies are tied to strongly flux-density-limited samples of radio galaxies with differing redshift distributions, which impose difficult selection biases for conclusive population studies (Hardcastle & Croston 2020). More sensitive surveys across a wider range of wavelengths have revealed a more morphologically diverse population of observed examples. These populations challenge existing paradigms, in particular those tying morphological FR classification to radio luminosity (e.g. Mingo et al. 2019). Furthermore, increased sensitivity in the radio region allows us to observe more galaxies, with some examples such as remnant galaxies (Murgia et al. 2011; Brienza et al. 2016a,b), which are thought to be

★ E-mail: inigo.slijepcevic@postgrad.manchester.ac.uk

FRI/FRII descendants, fitting into the existing paradigm. However, there are also examples that do not fit well into the FRI/II schema, such as hybrid radio galaxies (Gopal-Krishna & Wiita 2000). There is further ambiguity with a morphological split between compact and extended sources (Miraghaei & Best 2017); whether compact sources are a separate class of radio galaxy or an 'FR0' precursor to extended FRI/FRII sources is still unknown (Baldi, Capetti & Giovannini 2015). This raises the (open) question of whether FRI/FRII classification is an optimal classification scheme, particularly if we are using it as ground truth to train a supervised model.

When building samples of radio galaxies with which to test models linking morphology to physical characteristics, an additional complication is the volume of data produced by modern radio telescopes. The Rapid Australian Square Kilometre Array Pathfinder (ASKAP) Continuum Survey (McConnell et al. 2020) has already detected three million extended sources and the upcoming Evolutionary Map of the Universe (EMU) survey is expected to detect 70 million radio sources (Norris et al. 2006, 2011, 2021). Such large data volumes are not compatible with the *by eye* approaches to classification that have been widely used historically. This has resulted in an increased development and utilization of automated detection and classification. Consequently, machine learning methodologies have continued to gain traction as automated astronomical image classification tools. In particular, for future surveys with the Square Kilometre Array (SKA), which is expected to produce an unprecedented volume of data: ∼1 petabyte (PB) of image data per day (Hollitt et al. 2017), detecting up to 500 million new radio sources (Prandoni & Seymour 2015), machine learning approaches that can effectively process huge data volumes will become essential.

In particular, convolutional neural networks (CNNs) have been successfully applied to image-based classification of radio galaxies. The original work in this field is Aniyan & Thorat (2017) who classified radio galaxies into FRI- and FRII-type objects; other works have also incorporated object detection and classification (e.g. Wu et al. 2019; Wang et al. 2021), and attempts to use more novel techniques such as capsule networks (Lukic et al. 2018), attention gating (Bowles et al. 2021; Wang et al. 2021), and group-equivariant networks (Scaife & Porter 2021) to help improve performance and interpretability have also been implemented. Mohan et al. (2022) demonstrated that a variational-inference-based Bayesian deep-learning approach could be used to provide calibrated posterior uncertainties on individual radio galaxy classifications. Alternatives to CNN-based approaches for classification of radio galaxies include Ntwaetsile & Geach (2021) who extracted Haralick features from images as a rotationally invariant descriptor of radio galaxy morphology for input to a clustering analysis, and Sadeghi, Javaherian & Miraghaei (2021) who calculated image moments to implement a support vector machine approach to radio galaxy classification. Becker et al. (2021) provide a comprehensive survey of current supervised machine learning techniques.

In the context of upcoming radio surveys, it is important to know how much labelling is required to achieve good classification results. Because of the high cost of labelling, reducing the size of the required labelled data set is beneficial and will allow us to deploy working models earlier. We expect comparative model performance to be (mostly) constant across archival and new data sets: we can therefore test different algorithms and models on the MiraBest data set, and expect the results to hold for new surveys.

Currently, the archival data sets available for training radio galaxy classifiers are of comparable size to many of those used in computer vision (e.g. CIFAR; Krizhevsky 2009, with around 10^5 samples available). However, a fundamental difference is domain knowledge needed for creating labelled data sets, which has a much higher cost for radio astronomy data. As a result of this, *labels* are sparse in radio galaxy data sets, with labels only included for a small fraction of data. Labelled catalogues of radio galaxies contain of order 10^3 objects and the largest publicly available FR-labelled machine learning data set of radio galaxies is MiraBest (Miraghaei & Best 2017; Porter 2020), which has 1256 samples, orders of magnitude lower than the number of unlabelled images in its originating sky survey [Faint Images of the Radio Sky at Twenty centimeters (FIRST); Becker, White & Helfand 1995]. However, while labelled radio galaxy data sets are in the low-data regime, we note that they are still typically significantly larger than those associated with one- or few-shot learning; although the potential of such approaches to radio galaxy classification has been explored by Samudre et al. (2022).

Data augmentation using flips and rotations of the training data as originally proposed in Aniyan & Thorat (2017) for radio galaxy classification is widely used to mitigate overfitting due to this lack of (labelled) data. It has been shown that a well thought out augmentation strategy is crucial to performance in the low-data regime for radio galaxy classification (Maslej-Krešňáková, El Bouchefry & Butka 2021), and the potentially negative impact of unprincipled data augmentation, particularly in the case of Bayesian deep-learning approaches to radio galaxy classification, has been highlighted by Mohan et al. (2022). Therefore we note that regardless of the learning paradigm, augmentation strategy will remain an important variable in model training.

However, when adapting any computing technique for use in science, we must ask the question: is the algorithm we are using designed for our use case? How closely do our inputs match those used in the computer science literature?

It is clear that supervised learning is *not* designed for data sets limited by the number of available labels and a large proportion of (useful) unlabelled data, as these data samples are simply discarded. Therefore, in order to achieve optimal results there are two options: (i) label more data, or (ii) use an algorithm capable of leveraging the population of unlabelled samples. We focus on the second approach as option (i) will require more and more human labelling as data set and model sizes increase. Despite some success with crowd-sourcing labels (Banfield et al. 2015), manual labelling of radio galaxies through citizen science is unlikely to scale to the rate of data output from telescopes such as the SKA, as non-expert labelling usually requires large group consensus to accurately label samples. It therefore seems obvious that an algorithm able to incorporate information from *unlabelled* data into its predictions while still making maximal use of any available labels is desirable in the case of radio galaxy classification.

In this work, we investigate whether it is possible to achieve performance comparable to the current supervised state of the art but with many fewer labels by using semi-supervised learning (SSL). In doing so, we isolate and test the effect of real world complexities in the radio astronomy data such as non-identically distributed labelled and unlabelled data sets, unknown class imbalance in the unlabelled data, and the choice of labelled data set. Furthermore, we present some generalized diagnostics to predict the performance of SSL algorithms to a new problem in a computationally efficient manner.

The structure of this paper is as follows. In Section 2, we introduce the SSL paradigm, review its previous applications in astronomy, and highlight the challenges arising from data structure differences in this field. In Section 3, we introduce the data sets being used in this work and their processing. In Section 4, we introduce the SSL algorithm

being used in this work and describe its implementation and training.[1] In Section 5, we present the results of SSL for radio galaxy classification under various data conditions. In Section 6, we discuss these results in the context of the radio galaxy classification problem addressed in this work. In Section 7, we draw our conclusions.

2 SEMI-SUPERVISED LEARNING

Semi-supervised learning (SSL) refers to a set of solutions designed to leverage unlabelled data when a small labelled data set is available. A full theoretical overview of SSL is beyond the scope of this paper, but can be found in Chapelle, Scholkopf & Zien (2009). Unlabelled data can provide the classifier with extra information about the data manifold while also regularizing the model and mitigating overfitting in the low (labelled) data regime. The model is fed a set of image-label pairs $(\boldsymbol{x}_l, \boldsymbol{y}_l) \in \boldsymbol{X}_l$along with a (usually larger) set of unlabelled images $\boldsymbol{x}_u \in \boldsymbol{X}_u$. The goal is to predict labels for a set of reserved test samples, $\boldsymbol{x}_{test} \in \boldsymbol{X}_{test}$, or a new unseen data set.

SSL has an extensive literature with many different approaches that achieve good results. It is usually desirable for models to be consistent when making predictions on perturbations/transformations of the same image, which partly motivates the widespread use of data augmentation during supervised training (Chen, Dobriban & Lee 2019). This idea has also been adopted in SSL, with many algorithms implementing some form of *consistency regularization* that penalizes the classifier purely based on the consistency of its predictions given perturbations (augmentations) of the image, rather than the correctness of the prediction (which is unknown for unlabelled data). For example, Tarvainen & Valpola (2017) use an exponential moving average of previous model weights to enforce consistent predictions, while Miyato et al. (2019) take a different approach, using adversarial perturbations to force consistency in all directions around each data point, smoothing the decision manifold and pushing it away from the data. Alternatively, *pseudo-labelling* involves propagating 'soft' labels predicted by the model on to unlabelled data points when the model has high confidence (Dong-Hyun Lee 2013; Pham et al. 2020). The FixMatch algorithm (Sohn et al. 2020) combines consistency regularization and pseudo-labelling by propagating the pseudo-label of a weakly augmented image on to a strong augmentation of the same image. FixMatch is the focus of Section 4.

2.1 Applications of semi-supervised and self-supervised learning in astronomy

SSL has been used for a broad range of applications in astronomy. Examples include Marianer, Poznanski & Xavier Prochaska (2021) who focus on outlier detection in gravitational wave data, Richards et al. (2012) apply SSL to photometric supernova classification, and Hayat et al. (2021b) who improve state-of-the-art galactic distance estimations from images using SSL.

Perhaps the most relevant SSL applications for this work are in image classification, where the most popular approach has leveraged unsupervised models to learn a structured representation with unlabelled data first ('pre-training'), before fine-tuning (usually with labels) at the end, which is common in computer vision (Chen et al. 2019). For example, Ma et al. (2019) successfully trained an autoencoder to create a latent representation for radio galaxies and then fine-tuned this model in the standard supervised manner with all available labels and a cross-entropy loss, to improve model accuracy. Cutting edge contrastive learning methods have been used for the pre-training stage in galaxy morphology (Hayat et al. 2021a) and for gravitational lens/non-lens (Stein et al. 2021) classification, with promising results indicating that this technique may work in the radio astronomy domain as well.

Self-organizing maps have also been used to generate useful representations for radio data, from which classes can be inferred. This has been used successfully for identifying rare and unusual object morphologies (Ralph et al. 2019; Galvin et al. 2020). However, this technique is unable to use existing labels. Furthermore, experts are required to label a set of 'prototypical' samples generated by the model, which may be large in number or have mixed features. Generative adversarial networks (GANs; a self-supervised generative architecture) have been used for survey-to-survey image translation with some success (e.g. Schawinski et al. 2017). However, as described in Glaser et al. (2019), the technique is unlikely to work in the case of low- to high-resolution translation (so-called 'super-resolution') in the case of radio astronomy, due to the nature of interferometric measurements.

Unsupervised morphological classification is attempted in Spindler, Geach & Smith (2021) with a variational deep embedder. The data are clustered in the latent space of the model, which can be looked at by eye to interpret the morphological differences between clusters. While this is a worthwhile approach for separating data into meaningful groups, the cluster 'labels' (i.e. FRI/FRII classification scheme in our case) cannot be chosen a priori, which we may require if we want to use scientifically defined categories (e. g. FRI/FRII). Furthermore, there can be some intercluster overlap, and we cannot control which properties the model uses most of its capacity on, resulting in scientifically unimportant parts of the image (e. g. secondary sources) taking up much of the model's attention.

2.2 Data structure challenges

State-of-the-art SSL algorithms achieve impressive accuracies on standard benchmarking data sets with few labels: 97.13 per cent on CIFAR-10 (Krizhevsky 2009) with 4000 labels using LaplaceNet (Sellars, Aviles-Rivero & Schönlieb 2021), 97.64 per cent on SVHN (Netzer & Wang 2011) with 1000 labels using Meta Pseudo-Labels (Pham et al. 2020), and 94.83 per cent on STL-10 (Coates, Lee & Ng 2011) with FixMatch. However, less work has been done in assessing the robustness of SSL to real-world data sets, which may include unclean, covariate-shifted or prior-probability-shifted data, out-of-distribution unlabelled data, or even simply varying proportions of labelled/unlabelled data. Oliver et al. (2018) give a detailed analysis of the shortcomings of the SSL literature in the context of real-world applications.

In addition to the class ambiguity discussed in Section 1, astronomical data present different challenges to those explored in the SSL literature. In the SSL literature, it is widely assumed that $\boldsymbol{x}_{test}$, $\boldsymbol{x}_u$, and $\boldsymbol{x}_l$ are all drawn from the same distribution, as illustrated in Fig. 1(a). In the literature this holds as labels are typically discarded from a data set to mimic an unlabelled pool, ensuring that there is no difference other than size between $\boldsymbol{X}_u$ and $\boldsymbol{X}_l$. However, in astronomy and indeed many applications in observational science, $\boldsymbol{X}_u$ contains previously 'unseen' observations such that $\boldsymbol{X}_u$ and $\boldsymbol{X}_l$ are in general *not* identically distributed. This causes *covariate shift* between the unlabelled and labelled data. The distribution of labels $p_u(y)$ and $p_l(y)$ is also not identical in general, causing prior probability shift (Quiñonero-Candela et al. 2009). Depending on the specific differences in a given application, this can cause problems for SSL algorithms not designed to deal with this kind of data set shift.

[1] Code can be found at https://github.com/inigoval/fixmatch

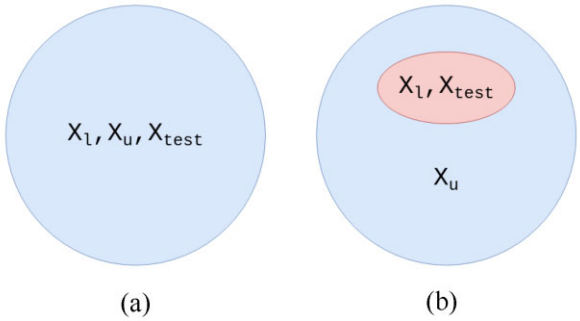


Figure 1. Visual representation of data structure differences between standardized machine learning benchmark data sets and real world radio astronomy data sets. (a) Data structure in the majority of the semi-supervised learning (SSL) literature, and (b) data structure in radio astronomy.

Of particular interest in astronomy is the problem of biased sampling of our training data. As a result of observational, instrumental, and intrinsic effects that skew towards selecting specific data such as particular ranges of flux density (brightness) and redshift (distance), the samples chosen for labelling are chosen in a biased way. While this is unavoidable when observing phenomena driven by complex natural processes, we must be wary of how it will affect our model performance. Observational data can also become corrupted due to image processing errors, for example during (de)compression (Ćiprijanović et al. 2021), which might lead to polluted unlabelled data. This makes it hard to how well our labelled data catalogues represent all observed data, with the consequent effects on model performance not always clear a priori. Specific examples of how this is being addressed for machine learning applications in astronomy include the use of Gaussian process modelling to improve data augmentation in photometric classification, making the training data more representative of test data (Boone 2019) and in galaxy merger classification where domain adaptation techniques have also been explored (Ćiprijanović et al. 2020). In both of these works, the authors are tackling data set shift between the labelled and test data.

Cai et al. (2021) use pre-training with the Swapping Assignments between Views (SwAV) algorithm (a type of contrastive learning; Caron et al. 2020) and fine-tuning with FixMatch (see Section 4 for details) to improve generalization on distribution-shifted unlabelled data. Their results imply that techniques using consistency regularization are more robust to data set shift than domain-adaptation-based approaches, and this motivates our choice of algorithm in this work.

3 DATA

We recreate the ideal SSL case shown in Fig. 1(a) to test whether SSL methods can help radio galaxy classification when little data are available and the data distribution is consistent between $\boldsymbol{X}_l$ and $\boldsymbol{X}_u$. To do this, we use a stratified subset of the MiraBest data set (Miraghaei & Best 2017; Porter 2020), as our possibly labelled data set $\boldsymbol{X}_l$, and discard labels from a proportion of the data set to create our unlabelled data set $\boldsymbol{X}_u$. For the more realistic case shown in Fig. 1(b), where our unlabelled data are drawn from a different distribution to our labelled data, we use data selected from the Radio Galaxy Zoo Data Release 1 (RGZ DR1) catalogue (Wong et al., in preparation) as $\boldsymbol{X}_u$.

Table 1. Three digit identifiers for sources in Miraghaei & Best (2017).

Digit 1	Digit 2	Digit 3
0: FRI	0: Confident	0: Standard
1: FRII	1: Uncertain	1: Double–Double
2: Hybrid		2: Wide Angle Tail
3: Unclassifiable		3: Diffuse
		4: Head–Tail

3.1 MiraBest

The MiraBest machine learning data set (Porter 2020) consists of 1256 images of radio galaxies pre-processed for deep learning tasks. The data set was constructed using the sample selection and classification described in Miraghaei & Best (2017), who made use of the parent galaxy sample from Best & Heckman (2012). Optical data from the Sloan Digital Sky Survey Data Release 7 (SDSS DR7; Abazajian et al. 2009) were cross-matched with NRAO VLA Sky Survey (NVSS; Condon et al. 1998) and FIRST (Becker et al. 1995) radio surveys. Parent galaxies were selected such that their radio counterparts had an AGN host rather than emission dominated by star formation. To enable classification of sources based on morphology, sources with multiple components in either of the radio catalogues were considered.

The morphological classification was done by visual inspection at three levels. (i) The sources were first classified as FRI/FRII based on the original classification scheme of Fanaroff & Riley (1974). Additionally, 35 'Hybrid' sources were identified as sources having FRI-like morphology on one side and FRII-like on the other. Of the 1329 extended sources inspected, 40 were determined to be unclassifiable. (ii) Each source was then flagged as 'Confident' or 'Uncertain' to represent the degree of belief in the human classification and, although this qualification was not extensively explained in the original paper, Mohan & Scaife (2021) have shown that it is correlated with model posterior variance over the data set. (iii) Some of the sources that did not fit exactly into the standard FRI/FRII dichotomy were given additional tags to identify their subtype. These subtypes include 53 'Wide Angle Tail' (WAT), nine 'Head–Tail' (HT), and five 'Double–Double' (DD) sources. To represent these three levels of classification, each source was given a three digit identifier as shown in Table 1.

To ensure the integrity of the machine learning data set, the following 73 objects out of the 1329 extended sources identified in the catalogue were not included: (i) 40 unclassifiable objects; (ii) 28 objects with extent greater than the chosen image size of 150 × 150 pixels; (iii) four objects that were found in overlapping regions of the FIRST survey; and (iv) one object in category 103 (FRII *Confident Diffuse*). Since this was the only instance of this category, it would not have been possible for the test set to be representative of the training set. The composition of the final data set is shown in Table 2. We do not include the subtypes in this table as we do not consider their classification in this work.

3.2 Radio Galaxy Zoo

The Radio Galaxy Zoo (RGZ) project is an online citizen science project that is aimed at the morphological classification of extended radio galaxies and the identification of host galaxies (Banfield et al. 2015). The online RGZ programme operated for ∼5.5 yr between 2013 December and 2019 May. Within this operational period, over 2.2 million independent classifications were registered for over

Table 2. MiraBest classwise composition.

Class	Confidence	No.
FRI	Confident	397
	Uncertain	194
FRII	Confident	436
	Uncertain	195
Hybrid	Confident	19
	Uncertain	15

140 000 subjects. The RGZ subjects inspected by citizen scientists consist of images derived from the FIRST survey (Becker et al. 1995) and the Asteroid Terrestrial-impact Last Alert System (ATLAS) survey (Norris et al. 2006), overlaid on to the *Wide-field Infrared Survey Explorer* (*WISE*) W1 (3.5 μm) infrared images. The RGZ data set used in this paper is from the RGZ DR1 catalogue (Wong et al., in preparation).

The RGZ DR1 catalogue contains approximately 100 000 source classifications with a user-weighted consensus fraction (consensus level) that is equal to or greater than 0.65. Within the catalogue, 99.2 per cent of classifications used radio data from the FIRST survey (Becker et al. 1995), and the remainder used data from the ATLAS survey (Norris et al. 2006). The largest angular size (LAS) of each source in the RGZ DR1 is estimated by measuring the hypotenuse of a rectangle that encompasses the entire radio source at the lowest radio contour (Banfield et al. 2015). This method is generally reliable if the radio components of a source are correctly identified and the source components are distributed in a linear structure (in projection). The RGZ DR1 catalogue is a catalogue of radio source classifications that matches up all related radio components to the host galaxy observed in the W1 infrared image. In this study, we used the catalogue of cross-matched host galaxies as our primary input sample.

3.3 Data formatting and cross-matching

Pre-processing was applied to both the MiraBest and RGZ DR1 data sets following the approach described in Aniyan & Thorat (2017) and Tang, Scaife & Leahy (2019). For the MiraBest data set this has been described previously in Bowles et al. (2021) and Scaife & Porter (2021). In this work, we create a machine learning data set from the RGZ DR1 catalogue with equivalent pre-processing.

For each object in the RGZ catalogue with $15 < \mathrm{LAS} < 270$ arcsec, the catalogued right ascension (RA) and declination (Dec.) were used to query the FIRST survey using the SkyView PYTHON API. From these search parameters the postage stamp server returns postage stamp images with 300 × 300 pixels. The FIRST pixel size is 1.8 arcsec.

The following pre-processing steps are then performed.

(i) Crop the central 150 × 150 pixels.

(ii) All pixels outside a radial distance of 0.6 times the LAS in pixels were set to zero.

(iii) All NaN-valued pixels were set to zero.

(iv) We use the 'outermost level' parameter from the RGZ DR1 catalogue to implement an amplitude thresholding. This is equivalent to $3 \times \sigma_{\mathrm{rms,\,local}}$ and we set all pixels with values below this level to zero.

(v) Finally, we normalize the image using

$$\mathrm{output} = 255 \times \frac{\mathrm{input} - \min(\mathrm{input})}{\max(\mathrm{input}) - \min(\mathrm{input})}, \tag{1}$$

where the minimum of the input is defined by the 'outermost level' parameter.

The final images with 150 × 150 pixels have a size equivalent to 270 arcsec (4.5 arcmin; 0.°075), which is the limiting LAS for selection from the RGZ catalogue.

RGZ DR1 entries were cross-matched with the catalogue of Miraghaei & Best (2017, MiraBest) using the VizieR PYTHON API. Based on the RA and Dec. of a source in RGZ DR1, matches were requested within a radius of 0.5 times the LAS, i. e. if an entry from the MiraBest catalogue was found within the unmasked field of view (FOV), then it was considered a match. The presence of a match (1) or no match (0) was recorded in the metadata for each data sample in the RGZ DR1 machine learning data set. Matched data samples were discarded from the data set in order to ensure that there was no overlap between the labelled MiraBest subset, $\boldsymbol{X}_{\mathrm{l}}$, and the unlabelled RGZ pool, $\boldsymbol{X}_{\mathrm{u}}$.

4 CHOICE OF SSL ALGORITHM

We selected the FixMatch algorithm (Sohn et al. 2020) from the pool of SSL techniques as it has been shown to achieve state-of-the-art performance on benchmarking data sets, has relatively few hyperparameters, and is computationally cheap. FixMatch has already been successfully applied across a broad range of domains, such as action recognition in videos (Singh et al. 2021), audio classification (Grollmisch & Cano 2021), and image classification (Sohn et al. 2020).

FixMatch makes use of unlabelled data through consistency regularization and pseudo-labelling by adding a loss term computed on two different augmentations of the same image. The 'weak' augmentation, denoted $\alpha(\cdot)$, retains the semantic meaning of the image and simply uses the typical rotations and flipping augmentations applied as standard in most deep learning classification implementations. The 'strong' augmentation, denoted by $\mathcal{A}(\cdot)$, may significantly alter the image, by applying a sequence of augmentations that do not necessarily preserve the semantic meaning of the image.

After augmentation, a *pseudo-label* is generated by assigning a label if the classifier softmax output is greater than a threshold, τ, for the weakly augmented image. This pseudo-label is then used as a target to compute the cross-entropy of the model's class prediction for the strongly augmented image and it is the corresponding 'pseudo'-cross-entropy that is minimized by the optimizer to train the model. The full form of the pseudo-loss term is given by

$$\mathcal{L}_{\mathrm{u}} = \sum_{u=0}^{\mu B} \underbrace{1(\max(p_{\mathrm{m}}(\alpha(\boldsymbol{x}_{\mathrm{u}}))) \geq \tau)}_{\text{threshold mask}} \times \underbrace{H(\hat{q}_{\mathrm{u}},\, p_{\mathrm{m}}(y|\mathcal{A}(\boldsymbol{x}_{\mathrm{u}})))}_{\text{pseudo-label cross-entropy}}, \tag{2}$$

where 1 is the indicator function and $\hat{q}_{\mathrm{u}}$ is the one-hot pseudo-label prediction on the weakly augmented image. B is the labelled batch size, μ and τ are hyperparameters controlling the unlabelled batch size and confidence threshold, respectively; p_{m} is the softmax output of the classifier. Fig. 2 illustrates graphically how an unlabelled data point flows through the model to update the gradients of the model.

The pseudo-loss term, $\mathcal{L}_{\mathrm{u}}$, is linearly combined with the supervised cross-entropy loss on the true labelled data using a tuning parameter,

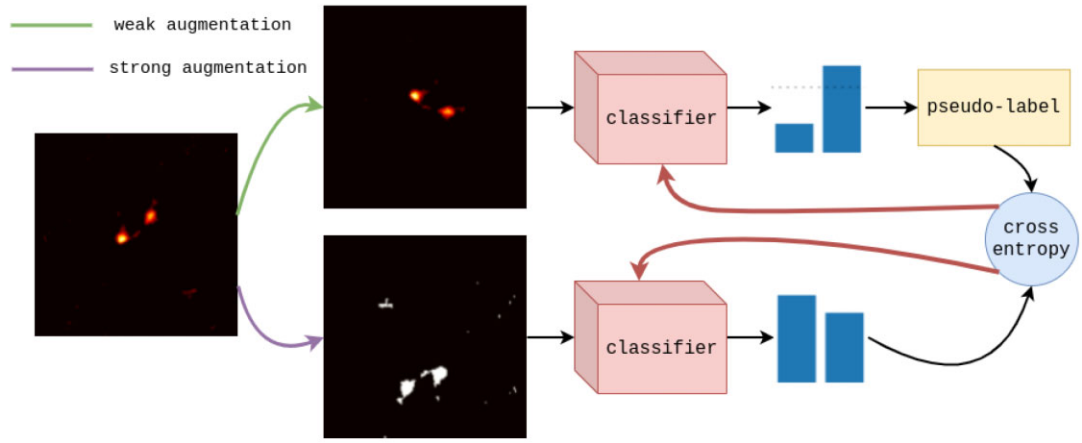


Figure 2. An unlabelled data point flowing through the FixMatch algorithm (Sohn et al. 2020). One strongly augmented image and one weakly augmented image are generated. The classifier makes a prediction on each, if the softmax value of the weakly augmented prediction is above a threshold value τ, consistency between the predictions on the two augmentations is enforced via the FixMatch loss.

λ_u, to give the complete FixMatch loss function:

$$\mathcal{L} = \lambda_u \mathcal{L}_u + \sum_{l=0}^{B} \underbrace{H(y_l,\, p_m(y|\alpha(\boldsymbol{x}_l)))}_{\text{supervised cross-entropy loss}}, \tag{3}$$

where in this work we use the canonical value of $\lambda_u = 1$.

4.1 Model architecture

To provide a realistic baseline, we use the convolutional architecture from Tang et al. (2019), which has high accuracy for radio galaxy classification and comparatively few (∼250 000) parameters that help avoid overfitting at low data volumes. The network has five convolutional layers with the Rectified Linear Unit (ReLU) activation and batch normalization for each layer, with three max pooling layers in total to reduce the dimensionality. This is followed by three fully connected layers, each with ReLU activation and a dropout layer. A softmax layer at the end squeezes predictions between 0 and 1. We refer to this architecture as the 'Tang network' (details can be found in Tang et al. 2019).

4.2 Data augmentations

4.2.1 Weak augmentations

We use the same weak augmentation for all algorithms. All pixel values in $\boldsymbol{X}_l$ and $\boldsymbol{X}_u$ are scaled to a mean of zero and a variance of one, calculated on $\boldsymbol{X}_l \cup \boldsymbol{X}_u$. Radio galaxy class is expected to be equivariant to orientation and chirality ('handedness'; see e. g. Ntwaetsile & Geach 2021; Scaife & Porter 2021) and we therefore rotate and flip each sample from $\boldsymbol{X}_l$ and $\boldsymbol{X}_u$ to a random orientation on ingest to the model.

4.2.2 Strong augmentations

We adjust the RandAugment algorithm (Cubuk et al. 2019) to generate strong augmentations that only include transformations valid on black and white images. RandAugment applies N sequential transformations with a given strength, m = randint(0, M), where M ∈ [0, 10] and larger values of m denote stronger distortions. Each sequential augmentation has a probability p_{aug} of occurring. We set $N = 2$, $M = 10$, $p_{aug} = 1$, consistent with Sohn et al. (2020).

Table 3 shows a list of the strong augmentations used in this work, all of which can be found in the PYTHON Imaging Library[2] or in our code. We note that although cropping is often used as a strong augmentation, we suggest that in this specific application cropping may be problematic as radio galaxy images are highly sparse (i.e. they contain a significant number of empty pixels), and that a randomly centred (upscaled) crop will mostly result in images of zeros. We found in our experiments that cropping significantly damages model performance and for this reason we omit it from the list of transformations passed to RandAugment.

4.3 Unresolved sources in the RGZ DR1 data set

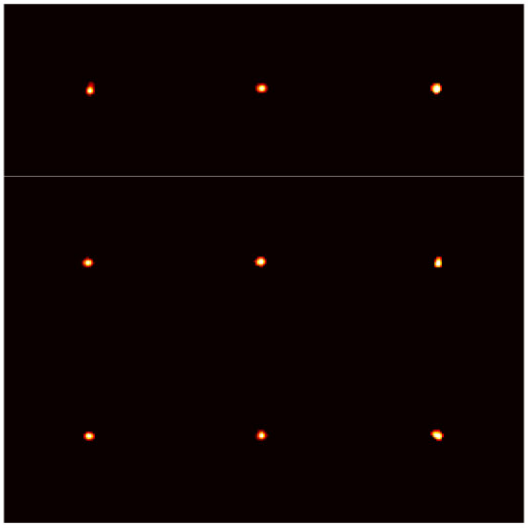
While the RGZ DR1 data originate from the same radio survey and are pre-processed in the same manner as MiraBest, the choice of filters for selecting data points in the RGZ DR1 case is significantly wider (∼10^5 data points) and its true class balance is unknown. Furthermore, we find that the RGZ DR1 data set contains many unresolved sources, see Fig. 3(a), which can overwhelm the learning signal. We remove these unresolved sources by enforcing a hard lower threshold on source extension. To calculate the exact value of this threshold, we estimate the similarity of the RGZ DR1 and MiraBest data sets over a range of threshold values and select the cut that gives the most similar data sets.

As a proxy for similarity we use a modification of the Fréchet inception distance (FID; Heusel et al. 2017) between MiraBest and RGZ DR1. The FID is a metric commonly used to evaluate the similarity of synthetic data sets to the training set for generative algorithms, such as when training GANs (Goodfellow et al. 2014), and has previously been adapted for use in radio astronomy when evaluating the synthetic radio galaxy populations generated from the latent space of a variational autoencoder (Bastien et al. 2021). To calculate this quantity, we first train the Tang network on the MiraBest data set using all available labels. We then separately forward pass both the RGZ DR1, denoted $\boldsymbol{X}_{RGZ}$, and MiraBest data set, denoted

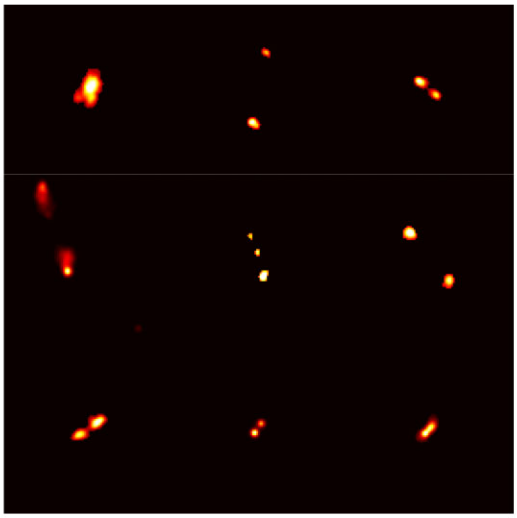
[2]https://pillow.readthedocs.io

Table 3. Strong augmentations used in this work. The table shows the name of each augmentation on the left, followed by a short description of the augmentation on the right.

Name	Description
AutoContrast	Cuts off some of the darkest/brightest pixels in the image histogram and then remaps the brightest/darkest pixels to white/black.
Brightness	Increases the pixel values of the image.
Contrast	Increase/decrease image contrast depending on input parameter.
Identity	No change to the image.
Sharpness	Blur/sharpen the image depending on input parameter.
Equalize	Forces grey-scale pixels into a uniform distribution.
Posterize	Reduces the number of possible magnitudes pixels can take by binning the pixel magnitudes.
ShearX	Distort the image in the x direction by shifting half of the image to the right and half to the left.
ShearY	Distort the image in the y direction by shifting half of the image upwards and half downwards.
Solarize	Invert all pixels above a given brightness threshold.
SolarizeAdd	Increase image brightness by a constant and then solarize.
TranslateX	Translate the image in the x direction.
TranslateY	Translate the image in the y direction.



(a) A random sample of unresolved sources from the RGZ data. Images are cropped to 95x95.



(b) A random sample of RGZ data with an angular size of 28 arsec and above. Images are cropped to 95x95.

Figure 3. Randomly selected samples from the RGZ DR1 data set at different cut thresholds.

$\boldsymbol{X}_{\rm MB}$, through the trained classifier, ignoring the classification outputs and instead taking the feature representations for each data set, denoted $\boldsymbol{F}_{\rm RGZ}$ and $\boldsymbol{F}_{\rm MB}$, respectively. The feature representations are simply the flattened output of the final convolutional layer of the Tang network (after activation, batch norm, and pooling). Fig. 4 illustrates this process. These feature representations can be interpreted as a low-dimensional embedding of each data set. We assume the features to follow a multidimensional Gaussian probability distribution for each data set, allowing us to calculate the mean, $\boldsymbol{\mu}$, and covariance matrix, $\mathbf{C}$, for both data sets. The Fréchet distance (FD) between the two data sets (in feature space) is then given by

$$\mathrm{FD}(\boldsymbol{X}_{\rm RGZ}, \boldsymbol{X}_{\rm MB}) = ||\boldsymbol{\mu}_{\rm RGZ} - \boldsymbol{\mu}_{\rm MB}||_2^2 + \mathrm{Tr}\left(\mathbf{C}_{\rm MB} + \mathbf{C}_{\rm RGZ} - 2(\mathbf{C}_{\rm RGZ}\mathbf{C}_{\rm MB})^{\frac{1}{2}}\right), \quad (4)$$

which gives us a similarity metric between the RGZ DR1 and MiraBest data sets, with lower FD values indicating a higher level of similarity.

In Fig. 5, it can be seen that there is an obvious minimum in the FD where the two data sets are most similar. As there is a small plateau in FD at the minimum, we use the lowest cut threshold that gives the minimum FD, equal to $\theta_{\rm cut} = 28$ arcsec, in order to remove the smallest number of samples. Examples of data samples from the MiraBest data set with angular sizes above this threshold are shown in Fig. 3(b).

4.4 Model training

We split the available data into validation and training sets by choosing randomly stratified splits (i. e. class balance is preserved). The training data are used to optimize the model weights and the validation set is used to optimize the hyperparameters of the model and to choose the best performing weights through a method known as early stopping.

We keep the validation set size realistic by scaling it to 20 per cent of $\boldsymbol{X}_1$ with a hard lower limit of 53 data points (5 per cent of the total MiraBest training data) to produce meaningful results when $\boldsymbol{X}_1$ is small, as an extremely small validation set causes the validation loss/accuracy to be too noisy, making it difficult to select a good model. We set the learning rate to 0.005 but scale the batch size, B, with $\boldsymbol{X}_1$. Astronomical data sets vary in size and content: large validation sets may not be available. Therefore, our models need to perform well across a range of scenarios without brittleness

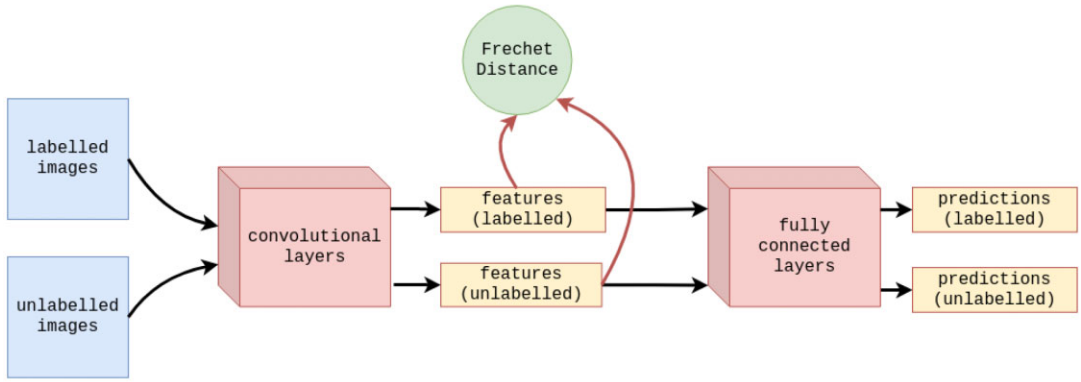


Figure 4. Schematic showing how to extract feature representations for Fréchet distance (FD) calculation. A supervised classifier is trained on all the labels. The (un)labelled data set is passed through the classifier and the output of the convolutional layers is eased as a representation for each data set. The similarity of the representations is computed using the FD.

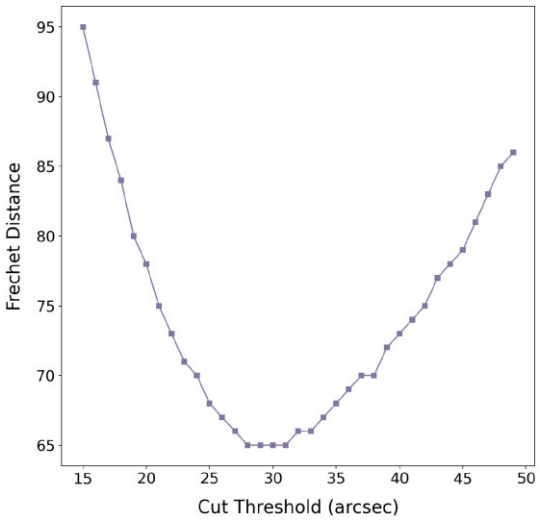


Figure 5. FD between the MiraBest and RGZ DR1 data sets as a function of different angular size cut thresholds, see Section 4.3 for details.

to small changes in hyperparameters. For these reasons, we keep hyperparameter tuning to a minimum, opting instead to choose reasonable values close to the optimal values in the computer vision literature (Sohn et al. 2020). We choose $\mu = 7$, which allows the model to use a large proportion of the unlabelled data in a single epoch, and $\tau = 0.95$, which allows the model's confidence on unlabelled data to pass the threshold before it begins to overfit. We allow a greater number of maximum training epochs when training using FixMatch as the model is less prone to overfitting and requires a greater number of epochs to result in softmax predictions above the threshold, τ, for the unlabelled data. Early stopping is used to choose the weights with the highest validation accuracy, which we empirically find to give better performance than using the validation loss.

Fair evaluation is a problem in the SSL literature (Oliver et al. 2018), yet is crucial if we wish to apply SSL methods to real data. To ensure reproducibility and to evaluate the variance in our results, each experiment is averaged over 10 runs initialized using the random seeds 0–9. This ensures consistent training and validation data splits during each experiment. A single test set is used to estimate the model's generalizability to unseen data after choosing the model weights with the best validation set accuracy.

Experiments were performed on a single Nvidia A100 GPU with a total of 28 d of runtime that includes all realizations of random seedings required for statistically significant results in some of the experiments. We used Weights & Biases (Biewald 2020) to track experiment results.

5 MODEL PERFORMANCE

5.1 Baseline results

To inform our decision making about analysis pipelines for future surveys, it is important to give a fair comparison with the current techniques used. This will help decision making when trading off increased complexity and computational overhead with performance for future pipeline design, as we expect our results on archival data to hold given new data sets in the same domain.

We establish a baseline model performance using the Tang network, see Section 4.1, by training in a fully supervised fashion with the data set weakly augmented, see Section 4.2.1, and the results of this baseline are shown in Table 4. As expected, it can be seen that baseline classification performance increases with the number of labelled samples. We see a steep drop off at low label volumes and a gentler increase at higher label volumes as the model approaches the accuracy with all labels.

Our baseline is in line with state-of-the-art performance in the literature for the MiraBest data set (including *Uncertain* samples), such as Scaife & Porter (2021) who achieve 85.30 ± 1.35 per cent accuracy by using an equivariant CNN, and Bowles et al. (2021) who achieve 84 per cent accuracy with an attention-gated CNN. We also note that the baseline achieves comparable accuracy to the fully labelled case with only 393 labels, indicating that the data set has significant redundancy in information.

5.2 Discarding labels to create an artificial SSL scenario (Case A)

In the SSL literature, algorithms are tested by throwing away labels from a labelled data set to create 'unlabelled' data. Although this approach is not realistic for a true use case in radio astronomy, it is important that we first isolate the effect of using an astronomical data set rather than a computer vision data set before testing on real

Table 4. Baseline FRI/FRII classification results on the MiraBest data set with different numbers of labels. Values given are means of 10 aggregated runs with uncertainties given by the standard error.

No. of labels	Accuracy (per cent)	Label	Precision	Recall	F1
10	60.9 ± 2.7	FRI	0.588 ± 0.027	0.647 ± 0.057	0.608 ± 0.035
		FRII	0.759 ± 0.018	0.731 ± 0.026	0.741 ± 0.014
50	76.2 ± 1.7	FRI	0.774 ± 0.025	0.728 ± 0.027	0.747 ± 0.020
		FRII	0.760 ± 0.016	0.794 ± 0.028	0.774 ± 0.017
101	81.6 ± 1.2	FRI	0.812 ± 0.020	0.818 ± 0.023	0.811 ± 0.012
		FRII	0.832 ± 0.016	0.815 ± 0.027	0.819 ± 0.014
203	84.1 ± 0.7	FRI	0.865 ± 0.018	0.804 ± 0.021	0.830 ± 0.008
		FRII	0.830 ± 0.014	0.876 ± 0.020	0.850 ± 0.008
301	84.4 ± 0.6	FRI	0.847 ± 0.001	0.830 ± 0.008	0.838 ± 0.006
		FRII	0.844 ± 0.006	0.858 ± 0.010	0.851 ± 0.005
393	86.4 ± 0.6	FRI	0.872 ± 0.007	0.843 ± 0.013	0.857 ± 0.007
		FRII	0.859 ± 0.010	0.884 ± 0.007	0.871 ± 0.006
855	86.9 ± 0.5	FRI	0.891 ± 0.004	0.831 ± 0.010	0.860 ± 0.007
		FRII	0.852 ± 0.008	0.905 ± 0.004	0.877 ± 0.005

unlabelled data. This gives us the most similar scenario to those explored in the SSL literature. To do this, we recreate the same artificial scenario by throwing away labels from the MiraBest data set. This allows us to set a best case upper bound on possible performance when using unlabelled data as we might naively expect the algorithm to perform worse or at best equivalently well to this scenario when using truly unlabelled data.

We use a small, randomly sampled subset of the MiraBest data set (Miraghaei & Best 2017) as $\boldsymbol{X}_{\rm l}$ and the remainder as $\boldsymbol{X}_{\rm u}$, equivalent to the case shown in Fig. 1(a). In all cases we keep the data stratified. Table 5 gives a comprehensive overview of our results. In Fig. 6, we plot the test-set accuracy and loss of Case A against the baseline model for different label volumes. It can be seen that the use of FixMatch in Case A achieves a consistently lower loss on the test data and outperforms the baseline, see Section 5.1, in test-set accuracy when there are 203 labels or fewer.

We are able to recover comparable accuracy (85.10 ± 1.13 per cent) to the supervised baseline with all labels (86.93 ± 0.54 per cent) using just 20 per cent (203) of the labels. However, FixMatch's performance degrades quickly with fewer than 50 labels, which is in stark contrast to similar experiments performed by Sohn et al. (2020) on more standard benchmark data sets, where good performance was achieved even with only one sample per class. However, we note that this 'sweet spot' where FixMatch has a significant advantage over the baseline model does not cover the full range of R (ratio of unlabelled to labelled data) and is quite narrow.

5.3 Testing FixMatch on real unlabelled data (Case B)

When training our model in a real scenario, we do not know what our unlabelled data contain. The model is unlikely to perform and in the 'ideal SSL' Case A scenario, so we test how close we can get when using real unlabelled data. This will give a better indication of whether the algorithm is robust enough for general use in radio astronomy, and if not, how much domain specific development is required. This will help decision making as to whether FixMatch is a useful stream of research for future pipeline development.

We train our model using FixMatch with a large pool of 20 000 unlabelled data samples from RGZ DR1, before applying a cut threshold as discussed in Section 4.3, with labelled samples from MiraBest. This is equivalent to the case shown in Fig. 1(b), where the unlabelled samples have some covariate shift from the labelled and test samples as a result of the selection biases inherent in the creation of the MiraBest catalogue.

In Fig. 6, we see an almost identical improvement (decrease) in test-set loss as in Case A (see Table 6 for all results). However, this does not translate to an improvement in accuracy – the model actually performs consistently *worse* than the baseline on our test set.

5.4 Model calibration

In order to perform science with our models, it is important that we can interpret our model output in a meaningful way. It is tempting to view softmax outputs as true probabilities, allowing us to find anomalous or out of distribution sources by simply looking at data with a low softmax output. However, the only constraint placed on the softmax outputs is that they must sum to 1 – there is no pressure from the loss function to calibrate softmax values to a probabilistic confidence. We can quantify how well calibrated the model is by using the expected calibration error (ECE; Guo et al. 2017), which we slightly modify.

We bin the test-set softmax outputs of the model into 10 bins, using a non-constant bin size such that we have the same number of data points in each bin. We calculate the error rate, κ_b, of the predictions (i. e. an estimate for the true probabilistic accuracy of the model) and the average of the softmax probabilities, μ_b, for each bin, $b \in B$. The ECE is then given by the average of these differences, with a lower ECE indicating better calibration:

$$\mathrm{ECE} = \sum_{b \in B} \frac{|\kappa_b - \mu_b|}{|B|}. \tag{5}$$

We calculate the ECE at different data volumes for the baseline, Case A and Case B. Fig. 7 shows our results, where we see a steep improvement in calibration as the number of samples initially increases for all algorithms. However, above ∼100 labels, the benefit of extra labels on calibration becomes much smaller. Furthermore, we do not see any benefit from using FixMatch on model calibration.

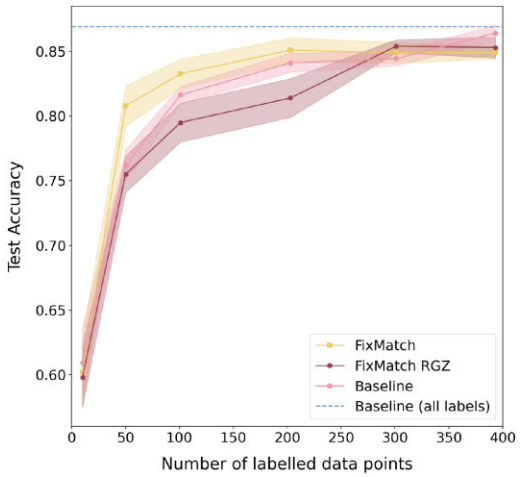
6 DISCUSSION

6.1 Does FixMatch outperform the baseline?

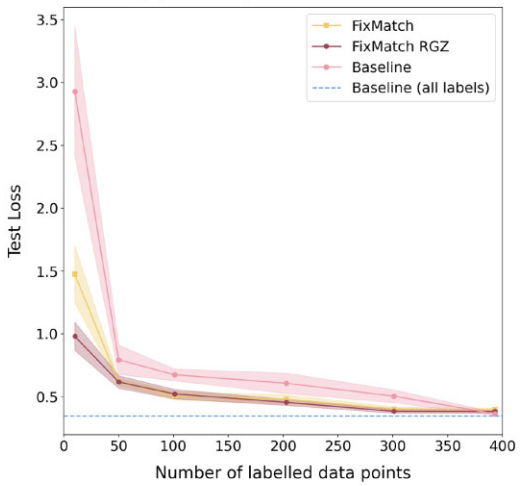
For our model to be useful, it needs to significantly outperform the baseline to justify the extra hyperparameters, complexity, and computational overhead. In Section 5.2, we see that FixMatch

Table 5. FixMatch FRI/FRII (Case A) classification results with different numbers of labels. Values given are means of 10 aggregated runs with uncertainties given by the standard error.

No. of labels	Accuracy (per cent)	Label	Precision	Recall	F1
10	60.2 ± 2.7	FRI	0.584 ± 0.029	0.703 ± 0.032	0.631 ± 0.021
		FRII	0.638 ± 0.024	0.5078 ± 0.060	0.551 ± 0.047
50	80.8 ± 1.6	FRI	0.803 ± 0.022	0.805 ± 0.022	0.802 ± 0.017
		FRII	0.819 ± 0.017	0.810 ± 0.024	0.812 ± 0.016
101	83.3 ± 1.3	FRI	0.824 ± 0.009	0.835 ± 0.015	0.828 ± 0.006
		FRII	0.846 ± 0.011	0.830 ± 0.013	0.837 ± 0.005
203	85.1 ± 1.1	FRI	0.869 ± 0.015	0.820 ± 0.028	0.840 ± 0.014
		FRII	0.845 ± 0.019	0.880 ± 0.018	0.859 ± 0.010
301	84.9 ± 0.8	FRI	0.853 ± 0.008	0.832 ± 0.016	0.842 ± 0.009
		FRII	0.848 ± 0.012	0.865 ± 0.009	0.856 ± 0.007
393	84.9 ± 0.5	FRI	0.869 ± 0.010	0.8122 ± 0.017	0.838 ± 0.007
		FRII	0.836 ± 0.011	0.884 ± 0.011	0.858 ± 0.004



(a) Accuracy on the test set.



(b) Loss on the test set.

Figure 6. Model performance as a function of labelled data set size for Case A, see Section 5.2 for details. Top: test accuracy. Bottom: test loss. R is the ratio of unlabelled to labelled data, and the performance of a fully supervised model is shown in each case as a dashed line.

consistently outperforms the baseline when training with 50–200 labels, achieving significantly higher test-set accuracy and lower loss. We hypothesize that this is both due to a regularization effect from the strongly augmented samples, as well as the model learning new information from the unlabelled data. This is illustrated in Fig. 8(a), where the well-behaved validation loss during training with FixMatch demonstrates its robustness to overfitting. We also observe that this effect is not as pronounced with 393 labelled samples, which is reflected in the equal minimum test loss in Fig. 8(b), implying that in the high data (label) limit, there is less benefit to using FixMatch.

In Section 5.3, we still see the same regularization effect as in Case A, as shown by the almost identical loss curves in Fig. 6(b) for both Case A and Case B. This suggests that the regularization effect of FixMatch is decoupled from the algorithm's ability to learn from the unlabelled data. However, Case B is unable to achieve better-than-baseline classification accuracy, regardless of the number of labels. We believe this is due to the covariate shift between the MiraBest data set, which comprises the labelled data and test set, and the unlabelled RGZ DR1 data set. The model may be learning to make better predictions on data from the RGZ DR1 distribution, whereas we are testing the model's performance on a test set drawn from the MiraBest data set, therefore our evaluation is skewed towards models that perform well on $\boldsymbol{X}_{\text{test}}$ rather than on truly unseen data. This means that although our model does not perform as well on our test set, it is unclear whether the opposite would be true if we our test set were drawn from the RGZ DR1 catalogue.

Since the two data sets are pre-processed identically and drawn from the same survey, the only other sources of covariate shift are the different filters and cuts used for each data set, see Section 3 for details. This can manifest itself in a number of ways, such as unseen subpopulations or class imbalance in the unlabelled data.

6.2 Effect of class imbalance in the unlabelled data set on model performance.

One possible cause of data set shift is label imbalance, more properly referred to as prior probability shift (Quiñonero-Candela et al. 2009). This is a commonly occurring problem in classification of astronomical data sets (e. g. photometric classification; Boone 2019). Since we cannot know the class balance of unlabelled astronomical data, it is important that model performance is not brittle to skewed unlabelled data class balance. Furthermore, this type of failure may be difficult to diagnose once the model has been deployed, as we

Table 6. FixMatch FRI/FRII (Case B) classification results with different numbers of labels. Values given are means of 10 aggregated runs with uncertainties given by the standard error.

No. of labels	Accuracy (per cent)	Label	Precision	Recall	F1
10	59.8 ± 2.3	FRI	0.580 ± 0.024	0.697 ± 0.048	0.622 ± 0.022
		FRII	0.644 ± 0.027	0.505 ± 0.061	0.545 ± 0.005
50	75.5 ± 1.4	FRI	0.760 ± 0.019	0.734 ± 0.035	0.740 ± 0.020
		FRII	0.764 ± 0.018	0.775 ± 0.030	0.764 ± 0.015
101	79.5 ± 1.5	FRI	0.792 ± 0.025	0.799 ± 0.019	0.791 ± 0.013
		FRII	0.810 ± 0.013	0.792 ± 0.031	0.798 ± 0.017
203	81.4 ± 1.5	FRI	0.825 ± 0.027	0.803 ± 0.026	0.807 ± 0.013
		FRII	0.825 ± 0.018	0.824 ± 0.041	0.816 ± 0.02
301	85.4 ± 0.5	FRI	0.876 ± 0.016	0.818 ± 0.015	0.843 ± 0.006
		FRII	0.841 ± 0.010	0.887 ± 0.018	0.862 ± 0.006
393	85.3 ± 0.8	FRI	0.865 ± 0.011	0.827 ± 0.018	0.844 ± 0.010
		FRII	0.846 ± 0.013	0.877 ± 0.011	0.861 ± 0.008

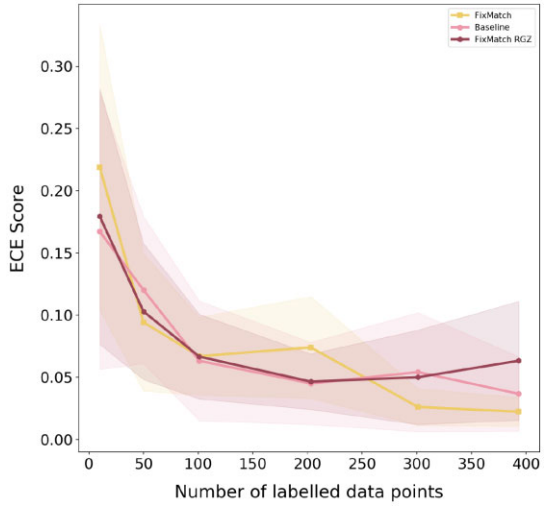
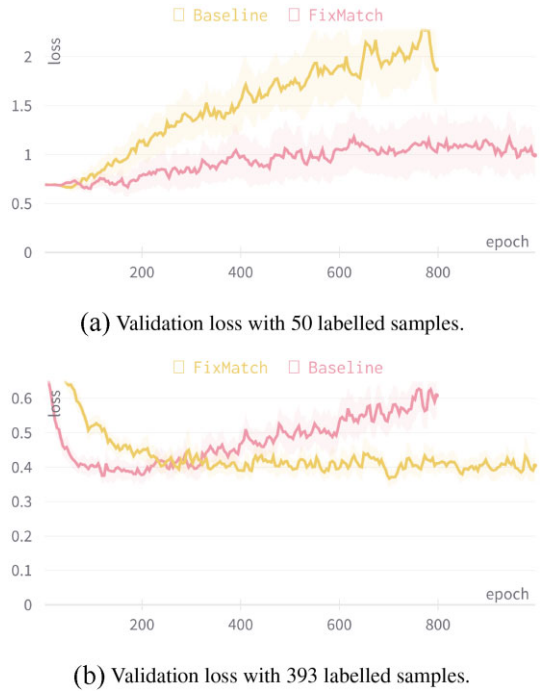


Figure 7. ECE scores with different numbers of labels. Values given are means of 10 aggregated runs with uncertainties given by the standard error.



(a) Validation loss with 50 labelled samples.

(b) Validation loss with 393 labelled samples.

Figure 8. FixMatch regularization effect on validation loss. The shaded area shows the standard error over 10 runs. We use exponential moving average smoothing with a weight of 0.3.

cannot check the unlabelled data for imbalance. However, while we cannot know the exact class balance of $\boldsymbol{X}_{\mathrm{u}}$, we can make an estimate by using an accurate classifier and looking at its predictions when tested on those unlabelled data. Fig. 9 shows the proportion of FRIs predicted by the baseline classifier trained using all labelled data and tested on the RGZ DR1 data set. The figure indicates these data likely have a high class imbalance regardless of cut threshold choice.

We can test whether the class imbalance in the RGZ DR1 data set is causing a difference in model performance due to prior probability shift by repeating Case A, but introducing an artificial imbalance into the unlabelled data pool and measuring the effect of this imbalance on test-set accuracy.

To examine whether this can fully explain the performance drop described in Section 5.3, we train a model on the MiraBest data set (Case A) with 69 labelled data samples and artificially imbalance the remaining unlabelled MiraBest samples by removing FRI/FRII samples until the unlabelled data set contains β FRI samples as a proportion of the total unlabelled samples. We then quantify label balance by computing $4(1-\beta)\beta$, which normalizes the value between 0 and 1 and makes it symmetric with respect to FRI or FRII imbalance. We perform this test using only the data qualified as *Confident* in the MiraBest data set in order to reduce noise in the results introduced by the higher predictive uncertainty associated with the *Uncertain* samples.

Fig. 10 shows the test-set accuracy as a function unlabelled class balance when removing either FRI or FRII samples from $\boldsymbol{X}_{\mathrm{u}}$. It can be seen that there is a clear trend in classifier test accuracy, which improves for more balanced $\boldsymbol{X}_{\mathrm{u}}$, indicating that class imbalance in $\boldsymbol{X}_{\mathrm{u}}$ is an important factor in the performance of the model when tested on $\boldsymbol{X}_{\mathrm{test}}$. Pearson R values of 0.914 and 0.852 show that the correlation between label balance and test-set accuracy is significant. These results suggest that class imbalance in the RGZ data set may be the cause of the poorer accuracy for Case B reported in Section 5.3.

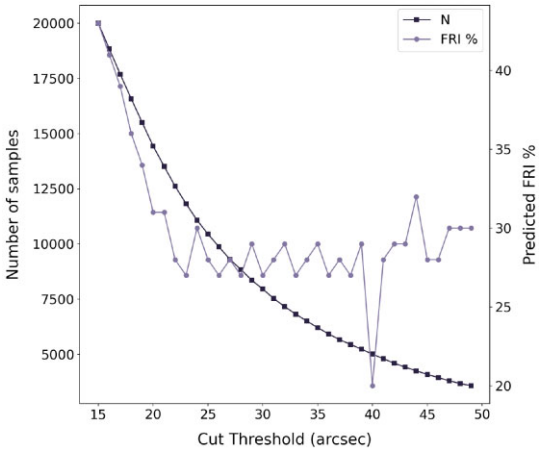
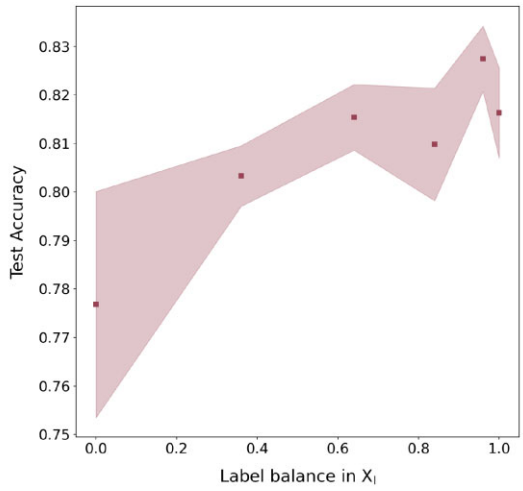
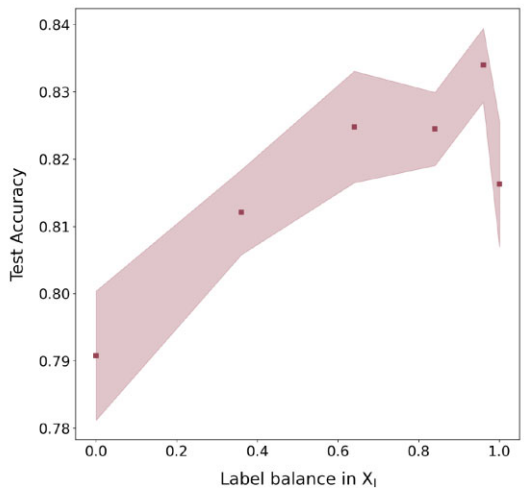


Figure 9. Number of samples remaining in the RGZ DR1 data set and predicted proportion of FRI-type objects as a function of angular size cut threshold.



(a) Accuracy on the test set after removing FRI samples from the unlabelled data-set. Pearson R coefficient: 0.914.



(b) Accuracy on the test set after removing FRII samples from the unlabelled data-set. Pearson R coefficient: 0.852.

Figure 10. Accuracy as a function of label balance in the unlabelled (MiraBest) data set. We only use *Confident* data in the labelled subset (69 labels) to reduce the noise in our results. Error bars show the standard error after aggregating 20 runs (seeded 0–19).

Table 7. Test-set accuracy using 50 labelled samples from different MiraBest categories (Case A, see Section 5.2 for details). The uncertainties given are the standard error calculated over 10 runs.

Algorithm	All (per cent)	Confident (per cent)	Uncertain (per cent)
Baseline	76.21 ± 1.67	81.90 ± 1.26	68.5 ± 2.13
FixMatch	80.78 ± 1.56	**82.29** ± 0.85	68.82 ± 2.56

6.3 Effect of labelled subset choice on model performance

During the experiments in Section 5.2, different data split, produced using from different random seeds, resulted in large discrepancies in test loss accuracy. For example, with 50 labelled samples there is a difference of ∼16 per cent between the best and worst runs, implying that some data split within the MiraBest data set are significantly more informative to the model than others. This may be due to the structure of the MiraBest data set, which is split into *Confident* and *Uncertain* categories, which indicate the human labeller's confidence in the label, see Section 3.1.

Understanding which labels are most informative to the model is an important goal, as it can inform which data samples to label to optimize the effectiveness of future models with the lowest labelling cost. To test this, we isolate training data sets from the *Confident* and *Uncertain* subpopulations to test which one of these is more informative to our model, or whether it is beneficial to use both.

We run experiments by training on 50 labelled samples from either the *Confident* subset, *Uncertain* subset, or a random sample of both (equivalent to Case A, see Section 5.2). The validation set is sampled from the same subset as the labelled data. In all cases, unlabelled data (if used) are randomly sampled from both *Confident* and *Uncertain* subsets. The results are shown in Table 7 where it can be seen that using only *Uncertain* labelled data greatly degrades model performance, decreasing accuracy by over 23 per cent. Using *Confident* samples alone gives the best performance, which is perhaps surprising given that the test set contains both *Confident* and *Uncertain* data samples, and may suggest that those samples with more ambiguous labelling, class overlap, or incorrect labels cause the classifier to learn a more disturbed decision function. This effect might potentially dominate any gains from the training samples spanning more of the data manifold, reducing accuracy of the final classifier. Given this result, we believe that the variance in performance seen when using both *Confident* and *Uncertain* is predominantly due to the variance in the number of *Confident* samples chosen for the labelled data set.

6.4 Can we use Fréchet distance as a proxy for model performance?

In general, testing models through repeated training and examination of test-set performance is computationally expensive. Furthermore, FixMatch has a significant computational overhead compared with the baseline, as many more data points need to be passed both forwards and backwards through the model. It would therefore be useful to have a predictor of the performance for the SSL model (FixMatch) that is fast to compute. This would allow us to efficiently test the viability of different unlabelled data sets.

One reason for poor SSL performance that is particularly relevant for our use case is covariate shift between the labelled and unlabelled data. Therefore, if we can measure the covariate shift in a computationally cheap manner, we may be able to predict whether an SSL technique is suitable for the problem. Here we do this using the Fréchet distance (FD) between $\boldsymbol{X}_u$ and $\boldsymbol{X}_l$, see Section 4.3. To

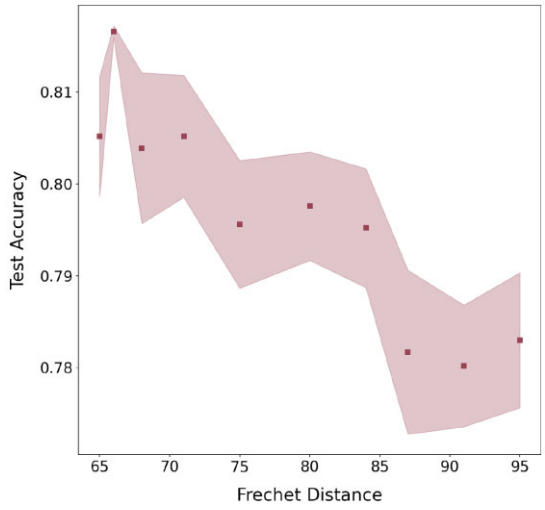


Figure 11. Test-set accuracy of FixMatch (Case B) as a function of FD between RGZ and MiraBest data sets. The shaded area shows the standard error calculated over 30 runs. Pearson R coefficient is 0.920.

test the effect of this covariate shift this we run Case B experiments and recover model performance as a function of cut threshold, which by proxy changes the FD, see Fig. 5, introducing a variable degree of covariate shift. For these tests we train on 65 data samples qualified as *Confident*, use the RGZ DR1 data set as the unlabelled data pool, and measure accuracy on the reserved test set. We only include data points to the left of the minimum (inclusive) in Fig. 5 as there is no motivation to remove more data than necessary from the unlabelled data.

The results of these tests shown in Fig. 11 show a negative correlation between the FD and the test accuracy (Pearson $R = 0.920$), indicating that we can use the FD as a predictor for model performance. However, we note that the results are quite noisy and in this case 30 runs were required to reduce the uncertainty bounds sufficiently enough for a significant result. Whilst the nature of this metric may seem a good proxy for covariate shift, the empirical nature of its calculation may cause it to be affected by other forms of sample-dependent data set shift, including prior probability shift, which may partly cause the noisy results. We therefore suggest that future applications of this metric would be strengthened by combining with a separate analysis to quantify the label imbalance in the unlabelled data that could give a decoupled measure of the prior probability shift.

7 CONCLUSION AND FUTURE WORK

In this work, we have demonstrated that the use of SSL, implemented using the FixMatch algorithm, provides some regularization benefits to radio galaxy classification when learning with few labelled samples, mitigating the effect of overfitting. Furthermore, we show that the model is able to learn from unlabelled data and that we achieve better accuracy on the test set using the SSL approach than the baseline with fewer labels in Case A. While this is encouraging, the improvement in accuracy is much smaller than for the standard benchmark computer vision data sets on which the algorithm was initially tested. We also find in Section 5.4 that using an SSL algorithm does not improve model calibration.

Poor results using the truly unlabelled RGZ DR1 data in Case B highlight an important obstacle to applying SSL 'in the wild' on scientific observational data where $\boldsymbol{X}_{\rm l}$ and $\boldsymbol{X}_{\rm u}$ are unlikely to be drawn from identical distributions. We find in Section 6.2 that class imbalance in the unlabelled data set is detrimental to classifier performance and this may account for the effect of prior probability shift between the unlabelled and labelled data on test-set accuracy. However, we also note that we are measuring performance on a test set that is distributed identically to the labelled data set, whereas real data may be distributed more similarly to the unlabelled data. Future work will look at estimating accuracy on unlabelled data, which would give a better estimate of model performance on real data. Furthermore, the effect of mismatched brightness distributions between the labelled and unlabelled data might be investigated. Since this semi-supervised scheme has not been designed for covariate shifted unlabelled data, we would expect that future improvements can push model performance towards the upper bound set by Case A (see Section 5.2 for details).

In Section 6.3, we see that the effect of a poorly chosen labelled data set on model performance is significantly greater than any improvement from the SSL algorithm. This highlights the importance of choosing to label data points that will be useful for our models. This result indicates that we need to be careful when choosing which data to label, as a poor choice can result in suboptimal performance. Active learning may be a suitable approach to begin solving this problem. It may also be useful for identifying samples with a high level of uncertainty during training, and mitigating the effect of propagating incorrect labels in these cases by correcting or removing these images.

In order to test our choice of unlabelled data set in a computationally efficient way, we use the FD to measure covariate shift between the unlabelled and labelled data. We find that this metric can predict test-set accuracy, and that it might be used as an initial guide when comparing different unlabelled data sets to narrow down a search. However, our results are noisy due to high sample variance and the metric does not directly measure label imbalance. We believe it would be best used as part of a fuller analysis that also takes into account prior probability shift (label imbalance).

We believe that a naive application of SSL, although it may outperform the baseline in some cases and provide useful regularization, requires further domain-specific development to provide a worthwhile advantage in the case of radio galaxy morphology classification. Although the FixMatch algorithm has minimal overhead when passing through labelled data, there is a significant computational cost in passing through the unlabelled data. Furthermore, future work will need to address the problem of biased sampling in our labelled data sets and the covariate shift between the unlabelled and labelled data set before the extra cost can be justified. There is also a limit on how much unlabelled data can be used in a semi-supervised way as we cannot keep increasing the μ parameter indefinitely. Unsupervised pre-training is one possible solution to these problems (Hendrycks et al. 2019), although this will require significantly more compute and larger models. We note that our work runs parallel to pre-training and that FixMatch can be used as a final fine-tuning layer, as has been done previously for computer vision applications (Cai et al. 2021; Kim et al. 2021). Furthermore, pre-training will allow larger volumes of unlabelled data to be used than is possible with FixMatch. StyleMatch (Zhou, Loy & Liu 2021) extends FixMatch to help with domain generalization, which may also allow data from different surveys to be used and could also be a useful area of research.

We suggest that radio astronomy specific development will likely be needed to achieve better results with both SSL and (contrastive) pre-training, as has been successfully done for gravitational lens identification (Stein et al. 2021). Specifically, a custom suite of augmentations for radio galaxy images should be developed, guided

by the computer science literature (Cubuk et al. 2018; Tian et al. 2020). While there have been attempts to automate this process (e.g. Tamkin, Wu & Goodman 2021), they are somewhat limited in the search space of possible augmentations, and do not include domain specific knowledge. We would therefore propose that augmentations are designed 'by hand' and tested for performance.

ACKNOWLEDGEMENTS

We thank the anonymous reviewer whose comments improved this work.

IVS, AMMS, MB, and MW gratefully acknowledge support from the UK Alan Turing Institute under grant reference EP/V030302/1. IVS gratefully acknowledges support from the Frankopan Foundation. HT gratefully acknowledges the support from the Shuimu Tsinghua Scholar Program of Tsinghua University.

This work has been made possible by the participation of more than 12 000 volunteers in the Radio Galaxy Zoo Project. The data in this paper are the result of the efforts of the Radio Galaxy Zoo volunteers, without whom none of this work would be possible. Their efforts are individually acknowledged at http://rgzauthors.galaxyzoo.org

IVS acknowledges the usefulness of https://github.com/kekmodel/FixMatch-pytorch for implementations.

DATA AVAILABILITY

Code for this paper can be found at https://github.com/inigoval/fixmatch

The RGZ DR1 catalogue will be made publicly available through Wong et al (in preparation). This work makes use of the MiraBest machine learning data set, which is publically available under a Creative Commons 4.0 license at https://doi.org/10.5281/zenodo.4288837

REFERENCES

Abazajian K. N. et al., 2009, ApJS, 182, 543
Aniyan A. K., Thorat K., 2017, ApJS, 230, 20
Baldi R. D., Capetti A., Giovannini G., 2015, A&A, 576, A38
Banfield J. K. et al., 2015, MNRAS, 453, 2326
Bastien D. J., Scaife A. M. M., Tang H., Bowles M., Porter F., 2021, MNRAS, 503, 3351
Becker R. H., White R. L., Helfand D. J., 1995, ApJ, 450, 559
Becker B., Vaccari M., Prescott M., Grobler T., 2021, MNRAS, 503, 1828
Best P. N., Heckman T. M., 2012, MNRAS, 421, 1569
Biewald L., 2020, Experiment Tracking with Weights and Biases. Available at: https://www.wandb.com/
Boone K., 2019, AJ, 158, 257
Bowles M., Scaife A. M., Porter F., Tang H., Bastien D. J., 2021, MNRAS, 501, 4579
Brienza M. et al., 2016a, A&A, 585, A29
Brienza M., Mahony E., Morganti R., Prandoni I., Godfrey L., 2016b, PoS, EXTRA-RADSUR2015, 068
Cai T., Gao R., Lee J. D., Lei Q., 2021, Proceedings of the 38th International Conference on Machine Learning, 139, 1170
Caron M., Misra I., Mairal J., Goyal P., Bojanowski P., Joulin A., Advances in Neural Information Processing Systems, 2020
Chapelle O., Scholkopf B., Zien A., eds, 2009, IEEE Trans. Neural Networks, 20, 542
Chen S., Dobriban E., Lee J. H., 2020, Advances in Neural Information Processing Systems
Ćiprijanović A., Kafkes D., Jenkins S., Downey K., Perdue G. N., Madireddy S., Johnston T., Nord B., 2020, Machine Learning and the Physical Sciences - Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS)
Ćiprijanović A. Kafkes D. Perdue G. et al., 2021, Fourth Workshop on Machine Learning and the Physical Sciences (35th Conference on Neural Information Processing Systems; NeurIPS2021)
Coates A., Lee H., Ng A. Y., 2011, J. Machine Learning Res., 15, 215
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
Cubuk E. D., Zoph B., Mane D., Vasudevan V., Le Q. V., 2018, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 113
Cubuk E. D., Zoph B., Shlens J., Le Q. V., 2020, IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 3008
Lee D.-H., 2013, in ICML 2013 Workshop: Challenges in Representation Learning
Fanaroff B. L., Riley J. M., 1974, MNRAS, 167, 31P
Galvin T. J. et al., 2020, MNRAS, 497, 2730
Glaser N., Wong O. I., Schawinski K., Zhang C., 2019, MNRAS, 487, 4190
Goodfellow I. J., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y., 2014, in Advances in Neural Information Processing Systems, preprint (arXiv:1406.2661)
Gopal-Krishna, Wiita P. J., 2000, A&A, 363, 507
Grollmisch S., Cano E., 2021, Electronics, 10, 1807
Guo C., Pleiss G., Sun Y., Weinberger K. Q., 2017, 34th International Conference on Machine Learning, ICML 2017, 3, 2130
Hardcastle M. J., 2018, MNRAS, 475, 2768
Hardcastle M. J., Croston J. H., 2020, New Astron. Rev., 88, 101539
Hayat M. A., Stein G., Harrington P., Lukić Z., Mustafa M., 2021, ApJ, 911, L33
Hayat M. A., Harrington P., Stein G., Lukić Z., Mustafa M., 2020, Third Workshop on Machine Learning and the Physical Sciences (35th Conference on Neural Information Processing Systems; NeurIPS2020)
Hendrycks D., Mazeika M., Kadavath S., Song D., 2019, Advances in Neural Information Processing Systems, 32
Heusel M., Ramsauer H., Unterthiner T., Nessler B., Hochreiter S., 2017, Advances in Neural Information Processing Systems, 6627
Hollitt C., Johnston-Hollitt M., Dehghan S., Frean M., Bulter-Yeoman T., 2017, in Lorente N. P. F., Shortridge K., Wayth R., eds, ASP Conf. Ser. Vol. 512, Astronomical Data Analysis Software and Systems XXV. Astron. Soc. Pac., San Francisco, p. 367
Ineson J., Croston J. H., Hardcastle M. J., Mingo B., 2017, MNRAS, 467, 1586
Kim B., Choo J., Kwon Y.-D., Joe S., Min S., Gwon Y., 2021, NeurIPS 2020 Workshop: Self-Supervised Learning - Theory and Practice
Krizhevsky A., 2009, Learning Multiple Layers of Features from Tiny Images. Technical Report, Science Department, University of Toronto
Lukic V., Bruggen M., Banfield J. K., Wong O. I., Rudnick L., Norris R. P., Simmons B., 2018, MNRAS, 476, 246
Ma Z., Zhu J., Zhu Y., Xu H., 2019, in Tan Y., Shi Y., eds, Data Mining and Big Data. Springer, Singapore, p. 191
McConnell D. et al., 2020, Publ. Astron. Soc. Aust., 37, e048
Marianer T., Poznanski D., Xavier Prochaska J., 2021, MNRAS, 500, 5408
Maslej-Krešnáková V., El Bouchefry K., Butka P., 2021, MNRAS, 505, 1464
Mguda Z., Faltenbacher A., Van der Heyden K., Gottlöber S., Cress C., Vaisanen P., Yepes G., 2015, MNRAS, 446, 3310
Mingo B. et al., 2019, MNRAS, 488, 2701
Miraghaei H., Best P. N., 2017, MNRAS, 466, 4346
Miyato T., Maeda S. I., Koyama M., Ishii S., 2019, IEEE Trans. Pattern Analysis Machine Intelligence, 41, 1979
Mohan D., Scaife A., 2021, Fourth Workshop on Machine Learning and the Physical Sciences (35th Conference on Neural Information Processing Systems; NeurIPS2021)
Mohan D., Scaife A. M. M., Porter F., Walmsley M., Bowles M., 2022, MNRAS, 511, 3722
Murgia M. et al., 2011, A&A, 526, A148
Netzer Y., Wang T., Coates A., Bissacco A., Wu B., Ng A.Y., 2011, in Neural Information Processing Systems. p. 1

Norris R. P. et al., 2006, AJ, 132, 2409
Norris R. P. et al., 2011, Publ. Astron. Soc. Aust., 28, 215
Norris R. P. et al., 2021, Publ. Astron. Soc. Aust., 38, 1
Ntwaetsile K., Geach J. E., 2021, MNRAS, 502, 3417
Oliver A., Odena A., Raffel C., Cubuk E. D., Goodfellow I. J., 2018, in 6th International Conference on Learning Representations, ICLR 2018 - Workshop Track Proceedings
Pham H., Dai Z., Xie Q., Luong M.-T., Le Q. V., 2021, IEEE Conference on Computer Vision and Pattern Recognition preprint (arXiv:2003.10580)
Porter F. A. M., 2020, MiraBest Batched Dataset (1.0) [Data set]. Zenodo, Available at: https://doi.org/10.5281/zenodo.4288837
Prandoni I., Seymour N., 2015, PoS, AASKA14, 067
Quiñonero-Candela J., Sugiyama M., Schwaighofer A., Lawrence N. D., eds, 2009, Dataset Shift in Machine Learning. The MIT Press, Cambridge, MA
Ralph N. O. et al., 2019, PASP, 131, 108011
Richards J. W., Homrighausen D., Freeman P. E., Schafer C. M., Poznanski D., 2012, MNRAS, 419, 1121
Sadeghi M., Javaherian M., Miraghaei H., 2021, AJ, 161, 94
Samudre A., George L. T., Bansal M., Wadadekar Y., 2021, MNRAS, 509, 2269
Saripalli L., 2012, AJ, 144, 85
Scaife A. M. M., Porter F., 2021, MNRAS, 503, 2369
Schawinski K., Zhang C., Zhang H., Fowler L., Santhanam G. K., 2017, MNRAS, 467, L110
Sellars P., Aviles-Rivero A. I., Schönlieb C.-B., 2021, preprint (arXiv:2106.04527)
Singh A., Chakraborty O., Varshney A., Panda R., Feris R., Saenko K., Das A., 2021, Computer Vision and Pattern Recognition, preprint (arXiv:2102.02751)
Sohn K. et al., 2020, Advances in Neural Information Processing Systems, preprint (arXiv:2001.07685)
Spindler A., Geach J. E., Smith M. J., 2021, MNRAS, 502, 985
Stein G., Harrington P., Blaum J., Medan T., Lukic Z., 2021, Fourth Workshop on Machine Learning and the Physical Sciences (NeurIPS 2021), preprint (arXiv:2110.13151)
Tamkin A., Wu M., Goodman N., 2021, International Conference on Learning Representations. preprint (arXiv:2010.07432)
Tang H., Scaife A. M., Leahy J. P., 2019, MNRAS, 488, 3358
Tarvainen A., Valpola H., 2017, Advances in Neural Information Processing Systems, preprint (arXiv:1703.01780)
Tian Y., Sun C., Poole B., Krishnan D., Schmid C., Isola P., 2020, Advances in Neural Information Processing Systems, preprint (arXiv:2005.10243)
Turner R. J., Shabala S. S., 2015, ApJ, 806, 59
Wang X., Wei J., Liu Y., Li J., Zhang Z., Chen J., Jiang B., 2021, Universe, 7, 211
Wu C. et al., 2019, MNRAS, 482, 1211
Zhou K., Loy C. C., Liu Z., 2021, NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications, preprint (arXiv:2106.00592)

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