



Attention-gating for improved radio galaxy classification

Micah Bowles^{1,2} Anna M. M. Scaife,^{1,2} Fiona Porter,¹ Hongming Tang^{1,2} and David J. Bastien^{1,3}

¹Jodrell Bank Centre for Astrophysics, Department of Physics & Astronomy, University of Manchester, Oxford Road, Manchester M13 9PL, UK

²The Alan Turing Institute, Euston Road, London NW1 2DB, UK

³Square Kilometre Array Organisation, Jodrell Bank Observatory, Macclesfield SK11 9FT, UK

Accepted 2020 December 18. Received 2020 December 18; in original form 2020 October 16

ABSTRACT

In this work, we introduce attention as a state-of-the-art mechanism for classification of radio galaxies, using convolutional neural networks. We present an attention-based model that performs on par with previous classifiers while using more than 50 per cent fewer parameters than the next smallest classic CNN application in this field. We demonstrate quantitatively how the selection of normalization and aggregation methods used in attention-gating can affect the output of individual models, and show that the resulting attention maps can be used to interpret the classification choices made by the model. We observe that the salient regions identified by our model align well with the regions an expert human classifier would attend to make equivalent classifications. We show that while the selection of normalization and aggregation may only minimally affect the performance of individual models, it can significantly affect the interpretability of the respective attention maps and by selecting a model which aligns well with how astronomers classify radio sources by eye, a user can employ the model in a more effective manner.

Key words: methods: statistical – techniques: image processing – radio continuum: galaxies.

1 INTRODUCTION

As astronomers collect larger and larger volumes of data, machine learning techniques are becoming increasingly prevalent in astronomical analysis. Examples include the use of Support Vector Machines (SVMs; see e.g. Cervantes et al. 2020) to study coronal mass ejections (Qu et al. 2006), gravitational lenses (Hartley et al. 2017), and for classifying astronomical objects detected in *GAIA* data release 2 (Gaia Collaboration et al. 2016; Bai, Liu & Wang 2018), and random forests (see e.g. Louppe 2014) for the classification of data from SDSS data release 15 (Aguado et al. 2019; Clarke et al. 2020). Neural networks in astronomy were reviewed as early as Miller (1993), and have been present ever since, for example in Lahav et al. (1996) for galaxy classification and as recently as Das & Sanders (2019) to estimate spectroscopic mass, age and distance for red giant stars.

In radio astronomy a massive increase in data volume is driving the adoption of machine learning methodologies and automation. This is due to the range of new instruments that have recently come online, including the Low-Frequency Array (LOFAR; Van Haarlem et al. 2013), the Murchison Widefield Array (MWA; Beardsley et al. 2019), the MeerKAT telescope (Taylor & Jarvis 2017), and the Australian SKA Pathfinder (ASKAP) telescope (Johnston et al. 2008). For these instruments a natural solution has been to automate the data-processing stages as much as possible, including classification of sources.

For example, the first fully public LOFAR Two-Metre Sky Survey (LOTSS; Shimwell et al. 2019) data release was mapped using a fully automated calibration and imaging process. This first data release covers only 2 per cent of the eventual coverage (424 square degrees)

and catalogues 325 649 sources. An object classification pipeline for LOTSS was developed and employed by Mingo et al. (2019), who identified 5805 classifiable sources with active galactic nuclei (AGNs) in LOTSS data release 1. This pipeline split the identified AGN into Fanaroff–Riley Class I (FRI) and Fanaroff–Riley Class II (FRII; Fanaroff & Riley 1974). This morphological classification of radio-loud AGNs separates ‘edge darkened’ FRI galaxies from ‘edge brightened’ FRII galaxies.

Over the years the morphological dichotomy of the FR classification has been widely investigated through both observations and simulations of complex sources (e.g. Mahatma et al. 2019; Smithard & Donohoe 2019; Schoenmakers et al. 2000). Various suggestions of additions and alternate classifications have also been made. A suggestion for the addition of an ‘FR0’ class was made in Baldi, Capetti & Giovannini (2015) for compact radio sources and a number of studies have investigated these sources, their properties and their relation to the prevailing FRI and FRII source classes (Baldi, Capetti & Giovannini 2016, 2019; Torresi et al. 2018; Capetti et al. 2020). From the LOTSS sample, Mingo et al. (2019) found that the traditional dichotomy of the FR classification was not sufficient to describe the complex morphologies of the detected sources, and they additionally classified ‘low-luminosity FRII’ sources, which extended to three orders of magnitude below the luminosity break observed to accompany the morphological distinction of FRI/II sources. Furthermore, various sub-classes of the FR scheme are commonly employed to label specific structures commonly found in radio galaxies. Both Miraghaci & Best (2017) and Mingo et al. (2019) include small samples of sources that do not clearly adhere to the FR dichotomy and it is certain that deeper and wider radio surveys will open up even more detailed views of such source morphologies. Additionally, a hybrid morphology, where a given source presents FRI properties on one side of its core and FRII properties on the other,

* E-mail: micah.r.bowles@gmail.com

Operations for Attention Gate (⊕)

- ▶ [1x1] Convolution
- ▶ ReLU
- ▶ Normalisation
- ▶ Upsampling
- ⊕ Addition
- ⊗ Multiplication

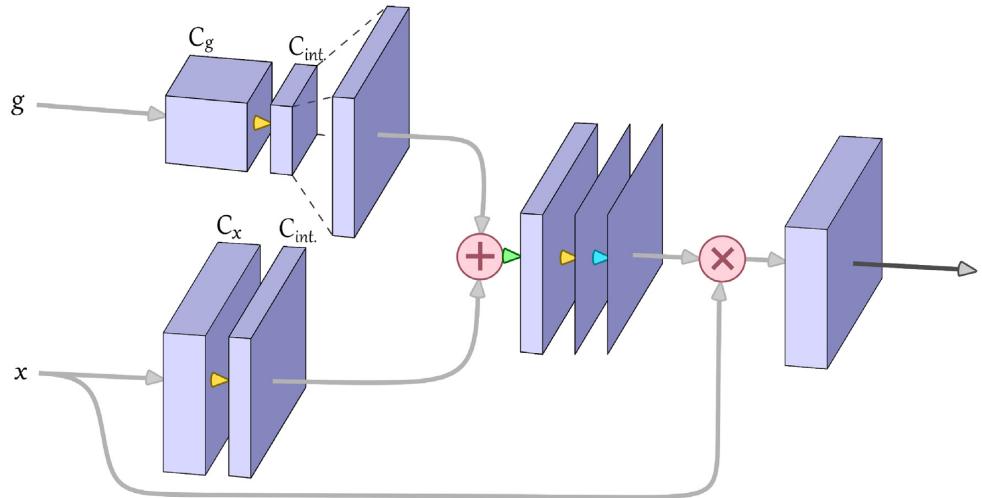


Figure 1. The attention gate implemented in this work, which learns soft feature and spatial visual attention (see equations 7 and 8).

Table 1. A summary of the primary network implemented in this work and visualized in Fig. 2, its parameters and the size of each of the feature maps. This is a valid for implementations with the fine-tuned aggregation, any normalization and three attention gates.

Operation Block	Output shape			Number of parameters
	Channel	Width	Height	
	1	150	150	0
[3 × 3] Conv. + ReLU + BN Norm	6	150	150	72
[3 × 3] Conv. + ReLU + BN Norm	6	150	150	342
[3 × 3] Conv. + ReLU + BN Norm	6	150	150	342
Max Pooling	6	75	75	0
[3 × 3] Conv. + ReLU + BN Norm	16	75	75	912
[3 × 3] Conv. + ReLU + BN Norm	16	75	75	2352
[3 × 3] Conv. + ReLU + BN Norm	16	75	75	2352
Max Pooling	16	37	37	0
[3 × 3] Conv. + ReLU + BN Norm	32	37	37	4704
[3 × 3] Conv. + ReLU + BN Norm	32	37	37	9312
[3 × 3] Conv. + ReLU + BN Norm	32	37	37	9312
Max Pooling	32	18	18	0
[3 × 3] Conv. + ReLU + BN Norm	64	18	18	18 624
[3 × 3] Conv. + ReLU + BN Norm	64	18	18	37 056
Attention gate 1	32	37	37	6209
Attention gate 2	16	75	75	5185
Attention gate 3	6	150	150	4545
Sum across height and width		54		0
Dropout		54		0
Aggregation function		2		14
Total parameters				101 447

is also widely recognized (Gopal-Krishna & Wiita 2000; Kapińska *et al.* 2017; Seymour *et al.* 2020).

In spite of these known issues, the FRI/II classification scheme persists, with recent research still seeking to understand through which processes and under what conditions these sources evolve (e.g. Ineson *et al.* 2015; Miraghaei & Best 2017). Large-scale samples of FRI/II classifications are used to evaluate theoretical and simulated population models (e.g. Godfrey, Morganti & Brienza 2017; Hardcastle 2018), model the evolution of radio-AGN (e.g. Best *et al.* 2014), and support the development of a unified AGN model (Netzer 2015). Further uses for these radio sources are summarized

by Hardcastle & Croston (2020). These include uses within cosmic magnetism (e.g. Bonafede *et al.* 2010; Govoni *et al.* 2010; O’Sullivan *et al.* 2019) and cosmology (e.g. Raccanelli *et al.* 2012). Regardless of the field using these sources, if the sources are not confidently classified, the application suffers.

It is therefore perhaps unsurprising that an increasing number of works in radio astronomy have been developing machine learning approaches to classify radio galaxies in the Fanaroff–Riley scheme (e.g. Aniyan & Thorat 2017; Lukic *et al.* 2019; Ma *et al.* 2019; Tang, Scaife & Leahy 2019). Since this is primarily a morphological classification, these approaches have focused on the use of convolutional

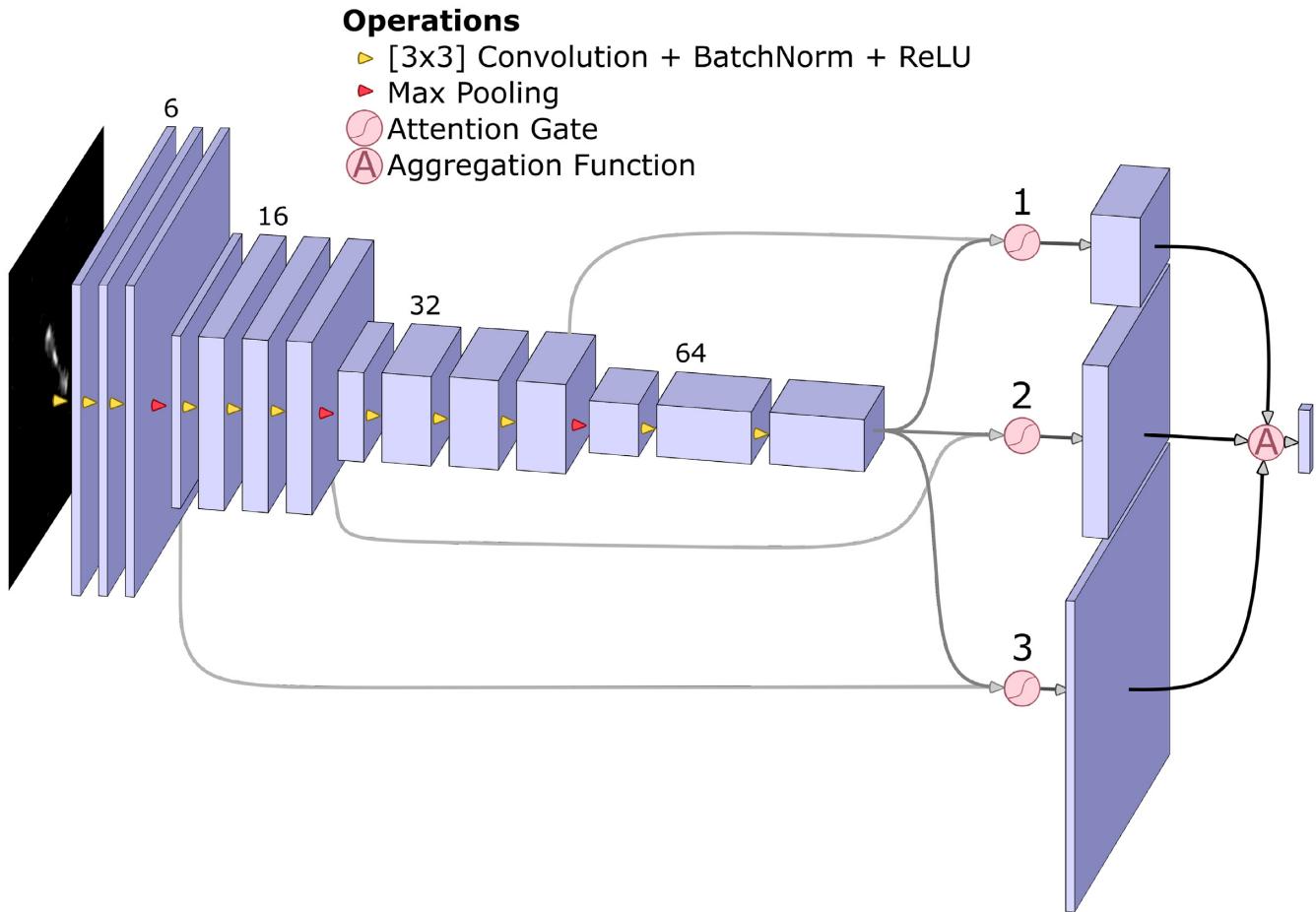


Figure 2. The primary network architecture implemented in this work. The attention gates are added in numbering order shown here, according to how many attention gates are requested. Channel width is given as a digit above each altered layer.

Table 2. Summary of attention gate 1, its parameters, and the size of each of the feature maps. These values are valid for any of the implemented normalizations as they have no learnable parameters.

Operation block	Output shape			Number of parameters
	Channel	Width	Height	
Input	32	37	37	9312
Global input	64	18	18	37 056
[3 × 3] Conv.	64	37	37	2048
[3 × 3] Conv.	64	18	18	4096
Upsample	64	37	37	0
[3 × 3] Conv.	1	37	37	65
Normalize att. map	1	37	37	0
Input × att. map	32	37	37	0
Attention gate 1				6209

Table 3. Attention normalization functions. Here, min and max are defined as functions that return the minimal and maximal values of their respective inputs, and μ and σ are the chosen mean and standard deviation of the input, which are selected to be $\mu = 0.5$ and $\sigma = 0.25$.

	Function	Functional form
(i)	Softmax	$\sigma_2(x) = e^x / \sum e^x$
(ii)	Sigmoid	$\sigma_2(x) = (1 + e^{-x})^{-1}$
(iii)	Range	$\sigma_2(x) = (x - \min(x)) / (\max(x) - \min(x))$
(iv)	Standardization	$\sigma_2(x) = (x - \mu) / \sigma$

neural networks (CNNs); however, the computational cost of training typical commercial CNNs with 10s to 100s of millions of parameters is not insignificant and furthermore, for problems with comparatively small training data sets such networks can easily result in over-fitting.

For example, the classic CNN architecture VGG16 (Simonyan & Zisserman 2014) is a deep convolutional neural network developed by the Visual Geometry Group at Oxford. VGG16 is able to fit complex visual tasks and can discriminate the 1000 classes of the ImageNet data set (Deng et al. 2009) with a top-1 value-error (i.e. the percentage of top predictions that are incorrect) of approximately 25 per cent. At the time of its publication, this was the highest performing model for the ImageNet data set and VGG16 became popular due to its structural simplicity and performance. However, with 138 M learnable parameters, VGG16 requires extremely large training data sets such as Imagenet, containing ~ 15 million images, to realize its purpose.

In radio astronomy, as with other domain specific applications of deep learning where labelled training data sets are often significantly smaller, it is beneficial to adjust the structure and size of CNNs. One such example is the network proposed and evaluated in Tang et al. (2019) which was trained on FRDEEP, a data set of 600 classified radio sources. With many fewer degrees of freedom, this problem does not require a network as deep as VGG16 and the implemented model has only 250 k parameters in comparison to the 138 M of VGG16.

Table 4. MiraBest data set summary. The original data set labels (MiraBest Label; Porter 2020) are shown in relation to the labels used in this work (Label). Hybrid sources are not included in this work, and therefore have no label assigned to them.

Label used	No.	Class	Confidence	Morphology	No.	MiraBest label
0	591	FRI	Certain	Standard	339	0
			Uncertain	Wide-angle tailed	49	1
			Uncertain	Head-tail	9	2
	631	FRII	Certain	Standard	191	3
			Uncertain	Wide-angle tailed	3	4
			Certain	Standard	432	5
NA	34	Hybrid	Uncertain	Double-double	4	6
			Certain	Standard	195	7
			Uncertain	NA	19	8
				NA	15	9

Similarly, the AG-Sononet architecture (Schlemper et al. 2018) was first introduced for classification in medical imaging, specifically to classify sonogram images. Although AG-SonoNet has the same number of convolution operations in its main body as VGG16, it does not use VGG16’s fully connected layers, which constitute 119 M (90 per cent) of that network’s 138 M parameters, and AG-SonoNet therefore contains only 696 k parameters. In order to achieve this reduction in parameter volume AG-SonoNet employs an attention gating mechanism to perform classification rather than a fully connected network.

Machine learning applications of attention are increasingly used to improve both the performance and interpretation of machine learning models (Ba, Mnih & Kavukcuoglu 2014; Stollenga et al. 2014; Bahdanau, Cho & Bengio 2015; Xu et al. 2015; Chen et al. 2017; Jetley et al. 2018). These applications are analogous to the biological concept of attention (Itti & Koch 2001; Zhou & Desimone 2011; Lindsay 2020), whereby the visual system prioritizes the most salient features in an image, i.e. the feature containing the most information pertinent to the context.

In this work, we introduce the concept of attention-gating to radio galaxy classification. We demonstrate that attention-gated networks can provide equivalent model performance to existing CNN-based radio galaxy classification whilst using significantly fewer trainable parameters. Furthermore, we demonstrate that the attention maps produced by these networks can be used to aid the interpretability of such machine learning models for astronomical applications. The structure of this paper is as follows: in Section 2, we introduce the attention mechanism for convolutional neural networks; in Section 3, we describe the network architecture deployed in this work and the implementation of the attention gates themselves; in Section 4, we give an overview of the radio astronomy data sets used for this work; in Section 5, we provide details of the model performance with reference to alternatives in the literature and justify our choice of normalization and aggregation method for the attention gates; in Section 6, we examine the average attention distribution as a function of target class across the data set and discuss its interpretation; in Section 7, we consider how attention distributions may inform a user about mis-classifications in a data set; and in Section 8, we summarize and draw our conclusions.

2 ATTENTION

There are two clear approaches to attention in machine learning: hard spatial attention and soft spatial feature attention. These two

approaches have clear alignments to overt and covert visual attention in the biological sense, respectively.

2.1 Hard versus soft spatial attention

When an ML algorithm outputs multiple sequential outputs based on individual sequential inputs selected by the model, this is considered hard attention. This has become common in natural language processing (Galassi, Lippi & Torroni 2019). For example, consider the input $\mathbf{x} = (x_1, \dots, x_T)^\top$, where each element of \mathbf{x} refers to an English word, which is to be translated to an output \mathbf{y} in another language. To do this, an encoder-decoder network is used where the encoder is a Bidirectional Recurrent Neural Network (BiRNN) and the decoder is comprised of an attention function and a Recurrent Neural Network (RNN). The BiRNN encodes the input to *annotations*, \mathbf{H} , as

$$\mathbf{H} = (\mathbf{h}_1, \dots, \mathbf{h}_T) = \text{BiRNN}(x_1, \dots, x_T). \quad (1)$$

The selected attention function uses these annotations to calculate a *context vector*, \mathbf{c}_t , for the next position t in the translated sentence as

$$\mathbf{c}_t = \text{Attention}(\mathbf{H}, \mathbf{s}_{t-1}), \quad (2)$$

where \mathbf{s}_i are hidden states used by the final RNN to extract the most probable next word in the translated sentence given the input and the translated sentence output so far:

$$p(y_t | y_1, \dots, y_{t-1}, \mathbf{x}) = \text{RNN}(\mathbf{c}_t, \mathbf{s}_{t-1}). \quad (3)$$

The attention operation itself is best explained in three steps: first, the annotations and hidden state of the previous output are passed into a function, f , which is typically a feedforward neural network (e.g. Bahdanau et al. 2015), which creates a set of scalar values, e_{ti} , that score the relatedness of the inputs around i to the outputs around t :

$$e_{ti} = f(\mathbf{s}_{t-1}, \mathbf{h}_i). \quad (4)$$

This value is normalized across each of the inputs, using a softmax function, to create a set of weights α_{ti} :

$$\alpha_{ti} = \frac{\exp e_{ti}}{\sum_{j=1}^T \exp e_{tj}}, \quad (5)$$

which are then used to scale the annotations and return the context vector:

$$\mathbf{c}_t = \sum_i \alpha_{ti} \mathbf{h}_i = \text{Attention}(\mathbf{H}, \mathbf{s}_{t-1}). \quad (6)$$

This kind of hard attention has also been used for computer vision applications such as generating image captions for a given input

by ‘translating’ image regions (Xu et al. 2015) and for classifying multiple objects in single images (Ba et al. 2014). Such hard attention selection for visual applications can be seen as being analogous to saccadic motion selecting salient regions with the fovea, and thus overt visual attention (e.g. Itti & Koch 2001).

Soft spatial attention, also referred to as feature attention, scales the representation according to spatial location equally across all feature maps and amplifies individual channels to scale certain features regardless of spatial location. Chen et al. (2017) used both soft spatial and feature attention for image captioning tasks, whereas Stollenga et al. (2014) generated soft feature attention by assigning weights to each feature map according to the feature maps generated in a forward pass of an image through the network.

The attention gates implemented in this work can be described as being *soft trainable attention*, which defines both spatial and feature saliency maps used to attend an input. These were initially introduced in Jetley et al. (2018), and were clarified and re-implemented in Schlemper et al. (2019).

2.2 Attention gates

Conceptually, attention gates amount to filters that prioritize salient features and their respective spatial locations within a given input. Soft feature attention is enforced using $[1 \times 1]$ -convolutions, which scale each feature (channel) of an input according to a learned weight. The attention weight is implemented as a normalized single-channel attention map which is a weighted sum of each of the salient features present in the image.

More specifically, an input, $\mathbf{x} \in \mathbb{R}^{C_x \times H_x \times W_x}$, is attended using a compatibility score, $C(\mathbf{x}, \mathbf{g}) : \mathbb{R}^{C_x \times H_x \times W_x} \times \mathbb{R}^{C_g \times H_g \times W_g} \rightarrow \mathbb{R}^{H_x \times W_x}$, to generate an (attended) output, $\alpha \in \mathbb{R}^{C_x \times H_x \times W_x}$, following:

$$\alpha_{ijk} = \sigma_2[C(\mathbf{x}, \mathbf{g})_{jk}] \cdot x_{ijk}, \quad (7)$$

with a normalization σ_2 and the compatibility score, $C(\mathbf{x}, \mathbf{g})$, calculated as

$$C(\mathbf{x}, \mathbf{g}) = [\sigma_1(\mathbf{x} * \theta + (\mathbf{g} * \Phi'))] * \Psi, \quad (8)$$

where the $*$ operator denotes the convolution operation; θ , ψ and Ψ are $[1 \times 1]$ convolutions; σ_1 is the ReLU non-linearity; and the prime in $(\mathbf{g} * \psi)'$ refers to the up-sampling required to match width and height dimensions to $(\mathbf{x} * \theta)$. This process is visualized in Fig. 1.

The $[1 \times 1]$ -convolutions are chosen such that $(\mathbf{x} * \theta)$ and $(\mathbf{g} * \psi)$ share the same intermediate channel number as their channel dimension output, which allows for simple addition of the two tensors. Furthermore, Ψ is a $[1 \times 1]$ -convolution that takes this intermediate channel dimension and reduces it to the one channel width of the compatibility score. Once normalized this compatibility score becomes the *attention map* (saliency map).

2.3 Aggregation methods

Once attention gates have been applied within a CNN, the different feature maps are used to generate an output. While in conventional CNN architectures this is achieved using multiple fully connected layers, in the case of attention-gated networks the number of fully connected layers is minimized in order to increase the classifier’s dependence on the attention gates themselves, the outputs from which are aggregated for classification.

Although in principle an aggregation method could be implemented in any way the user desires, in this work the four methods implemented by Schlemper et al. (2019) are considered. In these methods, the output of the attention gates remains as $\alpha^n \in \mathbb{R}^{C_n \times H_n \times W_n}$

with a superscript n to indicate which attention gate the output corresponds to. The output of each aggregation method is a vector with length equivalent to the number of classes $y \in \mathbb{R}^{n_{\text{classes}}}$, where in this work $n_{\text{classes}} = 2$. The output prediction for each class is evaluated as $y_{\text{out}} = \text{Softmax}(y)$.

The proposed mechanism classifies on the summed value of each channel of the attention gate’s output, meaning that the values f_i^n as defined by equation (9) are used to make the classification, such that

$$f_i^n = \sum_{j,k} \alpha_{ijk}^n, \quad (9)$$

with $i \in [0, \dots, C_n]$ and $n \in [1, \dots, N]$, where N is the selected number of attention gates. Logits are then constructed using one of the following four methods.

2.3.1 Mean

In this method the classification is made by taking the mean of multiple fully connected layers, each applied to the attention maps f_i^n :

$$y = \frac{1}{N} \sum_n \mathbf{W}_n f_i^n + \mathbf{b}_n, \quad (10)$$

where $\mathbf{W}_n \in \mathbb{R}^{2 \times C_n}$ and $\mathbf{b}_n \in \mathbb{R}^2$ are learnable parameters.

2.3.2 Concatenation

In this method, a classification is made on the concatenation of all of the feature maps:

$$y = \mathbf{W} \begin{pmatrix} f_i^1 \\ f_i^2 \\ f_i^3 \end{pmatrix} + \mathbf{b}, \quad (11)$$

where $\mathbf{W} \in \mathbb{R}^{2 \times (C_1 + C_2 + C_3)}$ and $\mathbf{b} \in \mathbb{R}^2$ are learnable parameters.

2.3.3 Deep supervised

The deep supervised method is an expanded version of the mean aggregation, where the final classification is an average of both individual classifications given in equation (10) and the concatenated classification given in equation (11):

$$y = \frac{1}{N+1} \left[\mathbf{W} \begin{pmatrix} f_i^1 \\ f_i^2 \\ f_i^3 \end{pmatrix} + \mathbf{b} + \sum_n \mathbf{W}_n f_i^n + \mathbf{b}_n \right], \quad (12)$$

where \mathbf{W} and \mathbf{b} are defined as in the definition of the concatenation method, and \mathbf{W}_n and \mathbf{b}_n are defined as in the definition of the mean method.

2.3.4 Fine tuned

The fine-tuned method employs a single fully connected layer to classify on the classifications made on each individual attention gate:

$$y = \mathbf{W}_{ft} \begin{pmatrix} \mathbf{W}_1 f_i^1 + \mathbf{b}_1 \\ \mathbf{W}_2 f_i^2 + \mathbf{b}_2 \\ \mathbf{W}_3 f_i^3 + \mathbf{b}_3 \end{pmatrix} + \mathbf{b}_{ft}, \quad (13)$$

where $\mathbf{W}_{ft} \in \mathbb{R}^{2 \times 6}$, $\mathbf{b}_{ft} \in \mathbb{R}^2$, $\mathbf{W}_n \in \mathbb{R}^{2 \times C_n}$, and $\mathbf{b}_n \in \mathbb{R}^2$ are all learnable parameters.

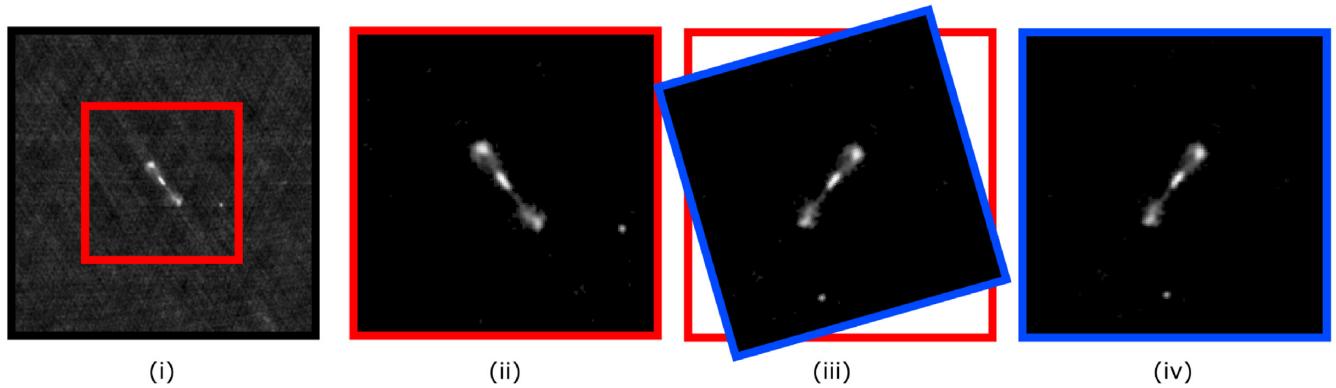


Figure 3. Illustration of the pre-processing of a source from the MiraBest test set. The individual panels present: (i) The original image extracted from the FIRST survey with a bounding box showing the applied crop. (ii) The cropped and sigma clipped image supplied by MiraBest. (iii) The equivalent data augmented image. (iv) A final augmented image with zero padding in remaining unfilled regions.

Table 5. Summary of the respective subsets of FR-DEEP and MiraBest, as implemented in this work. The Total* column refers to the totals augmented by 72 random transformations (the smallest number of random transformations applied). The training and validation data sets are split at random, and thus the values in the training and validation columns, which refer to specific (sub) classes, are approximate expectations rather than absolute values.

Data Set	Label	Total*	Total	Train	Validation	Test
FR-DEEP	FRI	19,008	264	193.6	48.4	22
	FRII	24,192	336	246.4	61.6	28
	Total	43,200	600	440	110	50
MiraBest	FRI	Certain	28,584	397	278	70
		Uncertain	13,968	194	135	34
	FRII	Certain	31,392	436	305	76
		Uncertain	14,040	195	137	34
Total (excl. hybrid)		87,984	1,222	855	213	153

Table 6. A comparison of a classical CNN classifier (from Tang et al. 2019) to an implemented attention gated network (with range normalization, aggregated using the fine-tuned method and three attention gates) trained and tested on the data set listed. *MiraBest** refers to the sub-set of MiraBest sources labelled as certain (see Section 4.1). The evaluation metrics are clarified in the Appendix.

Network Data set Class	Classic CNN FR-DEEP-F		AG-CNN FR-DEEP-F		AG-CNN MiraBest*		AG-CNN MiraBest	
	FRI	FRII	FRI	FRII	FRI	FRII	FRI	FRII
F1 score	0.90 ± 0.03	0.88 ± 0.06	0.87	0.90	0.91	0.92	0.82	0.86
Precision	0.95 ± 0.02	0.83 ± 0.04	0.87	0.90	0.89	0.89	0.91	0.80
Recall	0.85 ± 0.02	0.94 ± 0.04	0.87	0.90	0.95	0.94	0.75	0.93
Accuracy	$89 \pm 1\%$		88 %		92 %		84 %	
AUC	0.94		0.89		0.96		0.92	

3 NETWORK ARCHITECTURE

The network in this work is inspired by the SonoNet architecture (Schlemper et al. 2019), which in turn was inspired by the VGG-16 architecture (Simonyan & Zisserman 2014). We alter the base SonoNet architecture by removing the final pooling layer and all subsequent convolutional layers. This was done to reduce the complexity of the network and prevent overfitting. Overfitting is a serious problem when using deep networks and can be combated using validation methods. In this work, early stopping is implemented as the validation method of choice. As such a portion of the training data is reserved for validation. This model's loss on this validation set is recorded throughout training, with the final output being the model which achieved the minimal validation loss throughout training.

Validation tests showed that the original structure of the Sononet architecture applied to radio astronomy images quickly resulted in overfitting of the data and the network was truncated in response. Similarly, if there are not enough learnable parameters in a network, the trained model may not be able to differentiate between target classes and will either predict randomly or predict all sources to belong to a single class. For example, when the structure from Tang et al. (2019) was adapted to attention gates, this overfitting occurs immediately. The largest difference here, is that the fully connected layers, which contain 94 per cent of the original parameters, are removed for the attention gated implementation. Although the original performs well, the adapted model contains too few parameters and is not able to fit to the data correctly.

A detailed summary of the primary network implemented in this work and its parameters is given in Table 1 and depicted in Fig. 2. In the following sections, we refer to the architecture implemented here as the AG-CNN.

3.1 Attention implementation

As described in Section 2.2, the form of the attention gates used in this work is equivalent to that in Schlemper et al. (2019). However, the implementation of these gates within the network itself is not the same. In the architecture adopted here, the attention gates are implemented such that they are the only input on which the network makes a classification, thus guaranteeing that the attention gate outputs are used to make the final classification.

SonoNet uses two attention gates and additionally uses the final feature map of the convolutional layers as an input for the aggregation method. In principle, aggregation of these results allows a classification to be made based purely on the final feature map, and does not (theoretically) require information from the attention gates to be used at all. In this work, we implement up to three attention gates and only classify on the output of the attention gate(s) themselves. This is more similar to the initial learned CNN attention gate implementation (Jetley et al. 2018), where it was noted that classification under such a restriction is more consistent with the original concept of attention as practiced in NLP.

Table 2 summarizes the parameters of attention gate 1 as an example of how the parameters align within each attention gate (see also Fig. 1). In this work, as well as exploring each of the aggregation methods defined in Section 2.3, we also consider four different attention normalization functions for the attention gates, denoted σ_2 in equation (7). These are summarized in Table 3. Dropout is applied to the summed channel (feature) values output from the attention gates, prior to aggregation (see equation 9).

The network implementation is made in such a way that not all attention gate outputs are necessary for classification. Where classification is made using outputs from fewer than the maximum number of three attention gates the models include gates 1–3, as shown in Fig. 2, sequentially, i.e. 2 *attention gates* implies that gates 1 and 2 are used. In the case where no attention gates are included an additional max-pooling layer followed by a single fully connected layer is used for classification.

Models are trained over 100 epochs, using the Adam optimizer and an initial learning rate of 5×10^{-5} (adapted only once to enable the given model to train). The models trained on the MiraBest (Porter 2020) data set for this work took an average of 8 h 18 min \pm 1 h 20 min on an 8GB Nvidia RTX 2080 GPU. Depending on the hardware used and the optimization process, i.e. how many times the intermittent models are saved to disc, the training time can vary significantly. Further training time dependencies include the model's hyperparameters, e.g. the choice of optimizer, learning rate, training epochs, and data set. Given a pre-processed image, the trained model can output its predicted label in ~ 0.6 ms, as measured over 1000 augmentations of the MiraBest test set.

4 DATA

4.1 Data sets

The two data sets used in this work both use image data from the VLA FIRST survey (Becker, White & Helfand 1995), with the number (and sources) of labels derived from different sources. The data sets themselves are composed as follows.

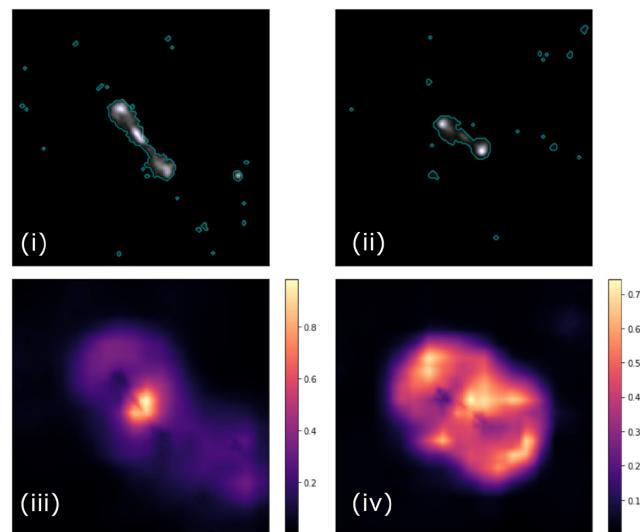


Figure 4. Example FRI and FRII sources selected for their respective clear classification with averaged attention maps derived from our model. (i) FRI example source: SDSS ID 2114-53848-625. (ii) FRII example source: SDSS ID 2266-53679-612. (iii) and (iv) are the averaged attention maps for the FRI and FRII sources, respectively.

Table 7. Evaluations of attention model with each of the four normalizations (fine-tuned aggregation and three attention gates).

Class	Normalization		Range norm.		Standardization		Sigmoid		Softmax	
	FRI	FRII	FRI	FRII	FRI	FRII	FRI	FRII	FRI	FRII
F1 Score	0.82	0.86	0.64	0.67	0.85	0.87	0.86	0.88		
Precision	0.91	0.80	0.65	0.66	0.90	0.84	0.92	0.84		
Recall	0.76	0.93	0.63	0.68	0.81	0.91	0.81	0.93		
Avg. accuracy	84 %		66 %		87 %		87 %			
AUC	0.92		0.66		0.92		0.93			

Table 8. Evaluations for each normalization and aggregation. The sets of models are averaged across aggregation to produce individual scores for the normalizations, and grouped by normalization to produce individual scores for the aggregations.

Norm. Class	Range norm.		Standardization		Sigmoid		Softmax	
	FRI	FRII	FRI	FRII	FRI	FRII	FRI	FRII
F1 Score	0.81	0.83	0.69	0.71	0.83	0.84	0.85	0.87
Precision	0.84	0.80	0.72	0.70	0.85	0.83	0.89	0.84
Recall	0.77	0.87	0.67	0.73	0.81	0.86	0.81	0.91
Accuracy	82 %		70 %		84 %		86 %	
AUC	0.87		0.71		0.85		0.92	
Agg.	Mean		Concatenation		Deep supervised		Fine tuned	
Class	FRI	FRII	FRI	FRII	FRI	FRII	FRI	FRII
F1 Score	0.78	0.82	0.82	0.84	0.79	0.78	0.79	0.82
Precision	0.85	0.77	0.86	0.81	0.77	0.80	0.83	0.79
Recall	0.72	0.88	0.78	0.88	0.81	0.75	0.76	0.85
Accuracy	80 %		83 %		78 %		81 %	
AUC	0.83		0.86		0.82		0.85	

4.1.1 FR-DEEP

The FR-DEEP data set was first presented in its entirety in Tang et al. (2019). The labels for the FR-DEEP data set were taken from the CONFIG (Gendre & Wall 2008; Gendre, Best & Wall 2010) and FRICAT (Capetti, Massaro & Baldi 2017) catalogues, where sources were visually classified by their expert authors. Tang et al. (2019)

selected a subset from those catalogues to include only sources that were denoted as confidently classified. In this work the FR-DEEP-F subset is used, which contains source images from the VLA FIRST survey at 1.4 GHz (Becker, White & Helfand 1995). The data set contains 264 images labelled FRI and 336 sources labelled FRII.

4.1.2 *MiraBest*

The MiraBest data set (Porter 2020) is comprised of 1256 radio galaxies with labels assigned using the catalogue of Miraghaei & Best (2017), where labels were assigned using visual inspection.

Although the MiraBest data set also contains sub-classifications of FR sources, in this work we use it only for binary FRI/FRII classification. Table 4 shows the labels used in this work in relation to the more detailed labels provided by the MiraBest data set itself. We do not include objects classified as *Hybrid*. Furthermore, the MiraBest data set flags individual objects as confidently classified (Certain) or unconfidently classified (Uncertain) depending on how much human interpretation was required to label a specific object, as described in Miraghaei & Best (2017). For the remainder of this work, we refer to the full Certain + Uncertain data set as *MiraBest* and to the Certain subset as *MiraBest**. Unless otherwise stated, the models in this work are trained and evaluated on MiraBest.

4.2 Pre-processing

FR-DEEP's pre-processing, as described in Tang et al. (2019), is the same pre-processing as was followed to create the MiraBest data set. The extracted images are processed in three stages before data augmentation is applied.

First the image is clipped: the image pixel values are set to zero if their value is **below a threshold of three times the root mean squared (rms) signal of the local noise** which was determined by a pixel histogram fit for each source image, this approach may clip out diffuse, low-surface brightness emission but was selected to align with Aniyan & Thorat (2017) who suggest this is the best clipping level for radio galaxy classification. This removes most artefacts and leaves behind cleaner images with clear sources. Any future inputs to the model should be treated in the same manner to allow the model to classify according to the features it has learned, and not the noise which the original image may contain.

The second step is to clip the image size to 150 by 150 pixels, i.e. 270 arcsec by 270 arcsec for FIRST, where each pixel corresponds to 1.8 arcsec by 1.8 arcsec. This is to standardize the size of the image and to provide the model with an image which ideally only contains the source of interest. However, by visual inspection we estimate that the clipped FR-DEEP-F and MiraBest data sets contain ~ 1.2 and 1.4 sources per image, respectively.

Finally, the image is then normalized as

$$\text{Final} = 255 \times \frac{\text{Img} - \min(\text{Img})}{\max(\text{Img}) - \min(\text{Img})}, \quad (14)$$

where 'Final' is the normalized image, 'Img' is the original image, and 'min' and 'max' are functions that return the single minimal and maximal values of their input, respectively. The steps in this pre-processing are illustrated in Fig. 3.

4.3 Data augmentation

Neither the FRDEEP nor MiraBest data sets are sufficiently abundant that all angles and possible positions of radio galaxies are represented within the data. Human observers easily recognize that there is no

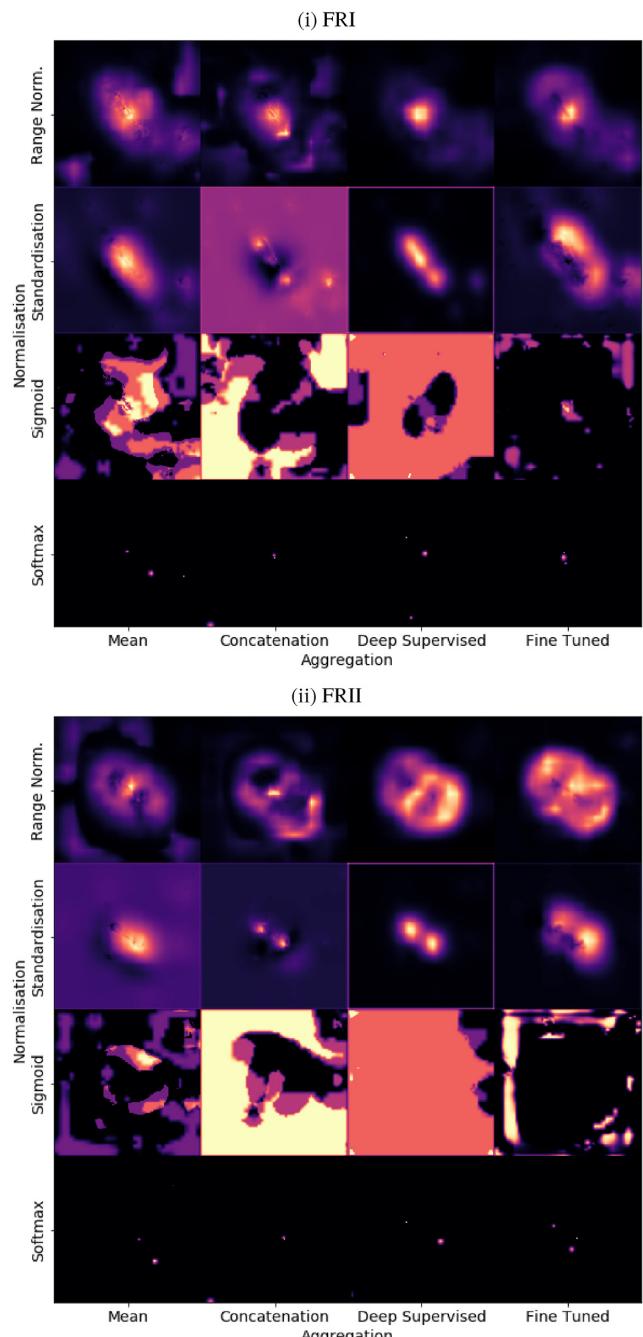


Figure 5. Array of averaged attention maps for (i) the FRI example source, and (ii) the FRII example source. The models each have three attention gates and differ by selection of normalization and aggregation methods.

class difference between an FRI galaxy and the same galaxy rotated by 180° ; however, the volume of data does not allow for the ML optimization to 'learn' this invariance. To allow for this, a commonly implemented process is *data augmentation*, which can be defined as the process by which the volume of training, validation and testing data is artificially inflated.

In the case of radio galaxies, for instance, the classification of a given source is independent of the scale, position or orientation of the sources within the image. Therefore, such invariances can be introduced during training by applying image transformations that enable the model to generalize to sources at scales, positions or

Table 9. Evaluations for models trained on MiraBest with varying number of attention gates (see Section 3.1 for clarification). The model with 1 attention gate was trained with a learning rate of 10^{-5} instead of the otherwise used 5×10^{-5} .

Gates	0		1		2		3	
Class	FRI	FRII	FRI	FRII	FRI	FRII	FRI	FRII
F1 Score	0.85	0.88	0.84	0.85	0.86	0.88	0.83	0.86
Precision	0.91	0.83	0.86	0.83	0.91	0.84	0.91	0.80
Recall	0.80	0.93	0.81	0.88	0.81	0.93	0.76	0.93
Accuracy	86 %		85 %		87 %		85 %	
AUC	0.93		0.90		0.94		0.92	

orientations it would otherwise be unfamiliar with. Fig. 3 shows the effects of the transformations used in this work. From the perspective of parameter optimization, the model is now able to sample the parameter space in positions which are valid, but would otherwise remain unsampled due to the limited size of the data set. We note that the transformations that are applied during data augmentation should not be selected without careful consideration of the problem itself as not all data sets are rotationally invariant.

The transformation function of a given input, $F(x)$, implemented in this work can be summarized as

$$F(x) = R_{[-180, 180]} \circ T_{[\pm 2px, \pm 2px]} \circ S_{[0.9, 1.1]}(x), \quad (15)$$

where $x \in [0, 1]^{150 \times 150}$ is the processed input image, S is a scaling translation that (from the centre of the image) randomly scales the input by a factor in the range $[0.9, 1.1]$, T is randomly applied translation operation that translates the image a number of pixels in $[-2, 2] \in \mathbb{N}$ both vertically and horizontally, where the pixel distance to translate is independently selected for each axis, i.e. $T_{[2, -1]}(x)$ is a valid translation, and R applies a rotation to the image around an angle randomly selected within the range $[-180, 180]$ using bilinear interpolation.

Assuming that at least three scales should be applied by S , and knowing that the transformations follow a uniform random selection, we estimate that the data can be augmented such that each image returns $360 \times 4^2 \times 3 = 17,280$ augmented images. In practice, the computational cost of parsing the full data set 17 280 times is extremely high. Consequently, to complete the training process within a reasonable amount of time, whilst maintaining the benefits of the data augmentation, the randomly transformed training data set is passed into the model 72 times during each epoch. This amounts to the model optimizing on 72 random augmentations of the training data set at each epoch. This means, that throughout training, the model will at most see 7200 separate data transformations of the estimated 17 280 possible transformations. During evaluation and testing, we use 360 random data set transformations, as the computational cost is comparatively negligible without the optimization stages.

The test sets are split from the original data sets, containing equivalent fractional populations of FRI and FRII sources, and are reserved for model evaluation. The remaining data samples are split randomly using an 80:20 training:validation split. The training set is used to optimize the network parameters, and the validation set is used for early stopping, i.e. the model that has the lowest validation loss throughout training is saved as the final model. Augmentation is applied after this split, ensuring that none of the unique sources from the test set appear within the training and/or validation sets.

Table 5 shows how the data sets are split for training, validation and testing, and the Total* column lists the augmented number of each data subset. This is done to provide an understanding of the scale of the relatively modest augmentation. It is important to note, that the

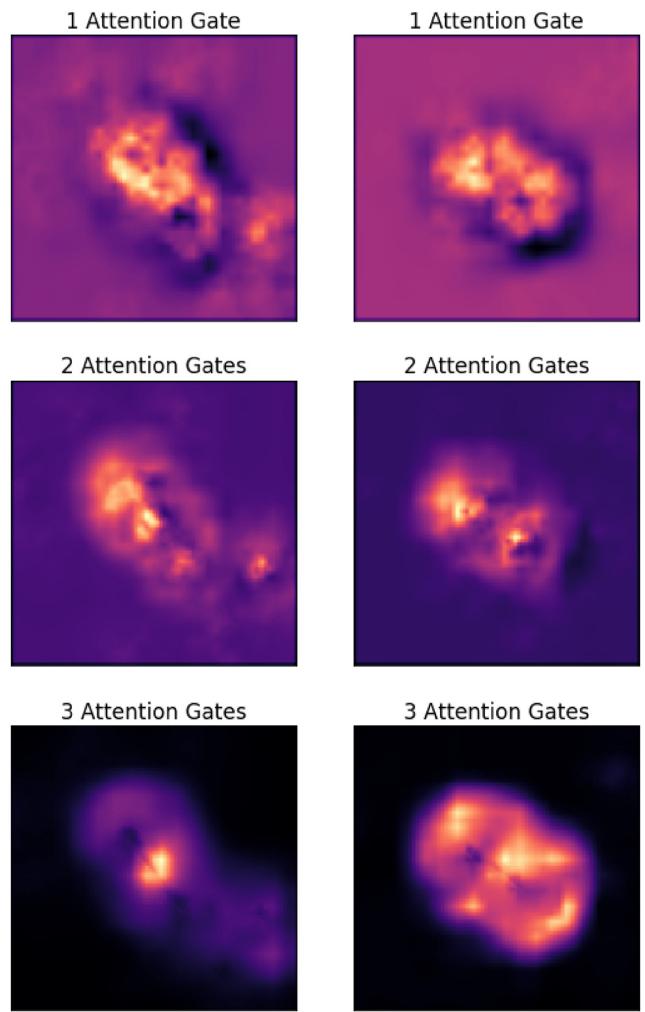


Figure 6. The (averaged) attentions of our models implemented with varying number of attention gates for the FRI (left) and FRII (right) example sources, as introduced in Fig. 4.

value of data augmentation is not a replacement for the inclusion of additional unique sources, or larger data sets, but rather is a supplementary process which allows data invariances to be exploited and reduce biases in the generalized model.

5 MODEL PERFORMANCE

To determine a baseline for our models on the MiraBest data set, we compare the performance of the AG-CNN architecture trained on various data sets to the *Classic CNN* model trained and evaluated in Tang et al. (2019). Table 6 lists the results of the evaluation of the various models.

From this table, it can be seen that the AG-CNN architecture performs similarly to the Classic CNN on the FR-DEEP data set, and in the case where it is trained on the MiraBest* data set the resulting model shows an improved performance. These observations are notable as the AG-CNN uses fewer than half as many parameters in comparison to the Classic CNN: 101 k compared with 250 k parameters (see Table 1). The relative performance loss seen for the model trained on the full MiraBest data set is expected, as the MiraBest-Uncertain subset of sources are more difficult to classify confidently for experts, and thus will cause larger error rates in the

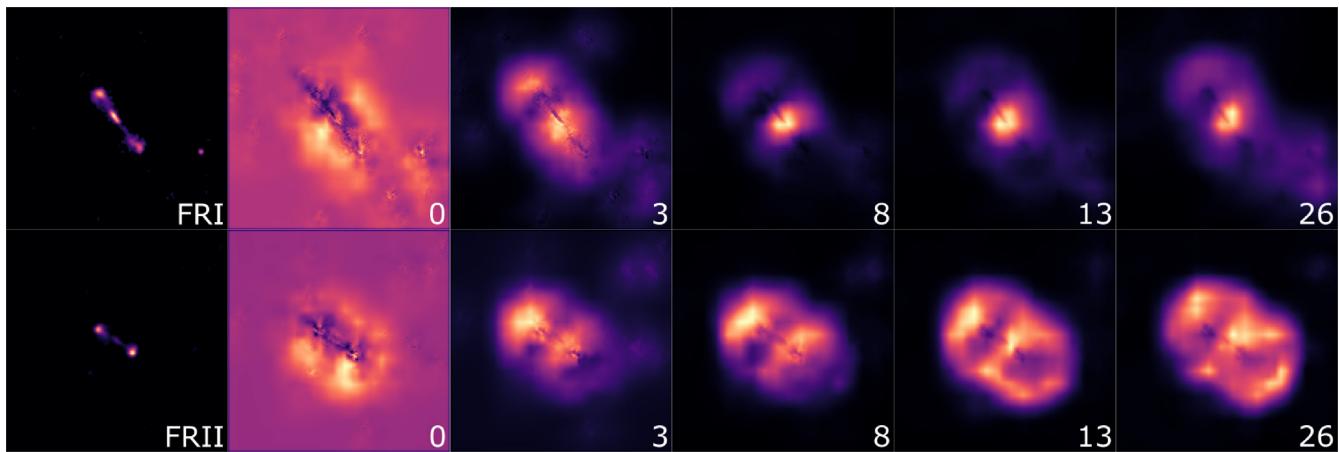


Figure 7. From left to right: Example sources from Fig. 4 followed by their attention maps at the respective epoch number throughout the training process. The epochs are sampled uniformly across the epochs where the model’s minimal validation loss was improved. Epoch 26 produced the model with the minimal validation loss, and thus the final model.

respective model, both because they contain edge cases that are more difficult to classify and because the labels of the sources themselves may not be fully correct.

Unlike traditional CNN models, the attention maps from the AG-CNN can also be visualized to understand how the model scales the features it extracts to make a final classification. To illustrate this, we use two example sources from the MiraBest data set. As an example of the FRI class, we use SDSS ID 2114-53848-625 ($J 13^{\text{h}}54^{\text{m}}33.0^{\text{s}}$ + $28^{\circ}14'36''$) and as an example of the FRII class we use SDSS ID 2266-53679-612 ($J 08^{\text{h}}04^{\text{m}}04.5^{\text{s}}$ + $15^{\circ}33'35''$). These sources and their combined attention maps are shown in Fig. 4. In this figure, it can be seen that both attention maps attend the region immediately around the source of interest, with the FRI example also attending the immediate vicinity of a secondary source to the bottom right of the image. While the FRI attention map primarily attends the central source, the FRII attention map primarily attends the immediate surroundings of the source, which is where the features relating to the lobes are expected to be found. Both images attend the region around the source more than the source itself, as shown by the dark regions in the shape of the sources extending through the middle of each attended region. We note the difference in scale between the averaged maps, with the FRI source peaking at 0.98 and the peak of the FRII region peaking at 0.74. This is a consequence of the range normalization used in the model and indicates that the peaks of the three individual attention maps align more closely in the FRI case, and less well in the FRII example.

5.1 Normalization and aggregation

To investigate how the selection of normalization function and aggregation function affect the model and its performance, we train multiple three gated attention models on the MiraBest data set. Table 7 shows how the model performs across each of the normalizations using the fine-tuned aggregation method and Table 8 shows how the model performs for each normalization averaged across all aggregation methods and for each aggregation method, averaged across normalization methods. These averaged evaluations show that the concatenation method has a slight performance gain over the other aggregation methods, and that softmax or sigmoid should be the primary choices for normalization based

on performance alone. However, whilst a model based purely on its performance may be of interest, for the attention mechanism to be maximally effective the interpretation of the attention maps themselves should also be taken into consideration.

We present in Fig. 5 the attention maps produced from different normalization and aggregation modes made using the example FRI/II sources that were presented in Fig. 4. From this figure, it is clear that the marginal improvements in performance seen for the softmax and sigmoid methods are gained at the expense of clarity and interpretability of the attention maps themselves. This result is consistent with the findings of Schlemper et al. (2019), who moved away from the use of softmax as a normalization method for attention gating due to the sparsity of the resulting attention maps.

5.2 Attention gate number

To investigate the impact of including different numbers of attention gates, we consider models trained on the MiraBest data set using the range normalization and fine-tuned aggregation methods with a varying number of attention gates. Table 9 displays the resulting evaluations of the respective models.

In this case, the highest performing model in terms of accuracy is the model with two attention gates. However, as with the investigation of aggregation and normalization methods, the resulting attention maps also play a part in evaluating the value of a given model. Fig. 6 visualizes the average attention maps when considering the example FRI and FRII sources. Here, one can see that the saliency achieved by the models with 1 and 2 attention gates is more dispersed. Although this does not hinder the model’s performance, it is potentially more difficult for a human observer to interpret.

Schlemper et al. (2019) state that they empirically found a third attention gate did not provide additional value to the system, but Jetley et al. (2018) state that the third attention gate encourages the model to learn salient features earlier in the network, as these features are used to make a third of the classification since they used mean aggregation. We suggest that the selection of attention gates should be considered with reference to the specific data problem and that the ability of the user to relate to the resulting attention distributions should be considered a factor in this process.

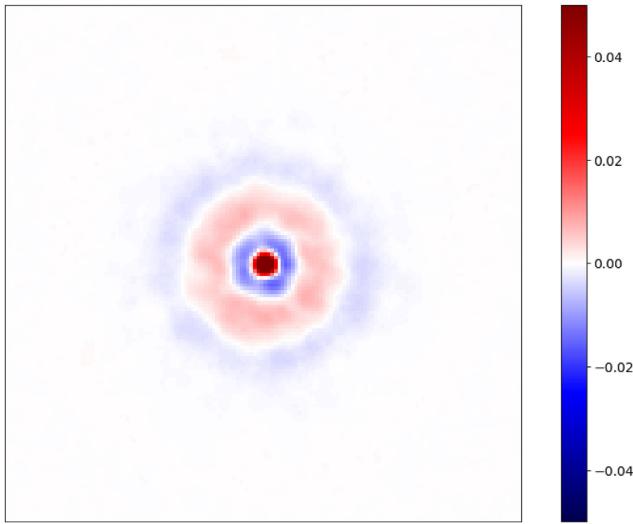


Figure 8. The difference between the mean pixel values of the 100 fold augmented FRI and FRII MiraBest test sets, $\langle \text{FRI} \rangle - \langle \text{FRII} \rangle$. To emphasize the structures in this difference image it is saturated at ± 0.05 , with an original maximal difference of 0.1 at the centre.

5.3 Attention as a function of epoch

Fig. 7 shows the two example sources and their attention maps at different points throughout the training process. As a function of epoch, the attention develops to become more specific to the regions discussed previously: the attention maps show no clear feature selection at epoch 0, but become more specialized throughout training. Note that in the FRI case, after the model learns to attend the central source (epoch 8) it shifts to attend the region around the source to a higher degree than it had previously. For the FRII case, the attended regions of epochs 3 and 8 are distributed asymmetrically around the source, with a larger focus on an offset region above the source’s lobes. By epoch 26, the learned attention is much more symmetrical for this example.

6 AGGREGATE ATTENTION

While presenting the attention maps for example sources is helpful when considering the model in individual cases, we must also consider the population average of attention across the data set. We calculate these averaged images by considering the mean of a randomly transformed set of images: both the test set itself as well as the attention maps derived from the augmented test set. Each attention map in this section is created by taking the pixel mean across a set of attention maps generated by passing the augmented test set into the model 100 times. We note that the statistics of the attention distribution across the augmentations is not Gaussian and will be considered in more detail in future work. As such, the mean is used as a comparative measure rather than a true parametrization.

For the input images themselves, Fig. 8 presents the difference between the mean pixel intensities of the FRI and FRII classes. Although there is clear structure, the individual sources are far less well defined than this figure may suggest, and we therefore caution against overinterpretation. However, a broad interpretation suggests that the central pixels are dominated by FRI sources, as one would expect from the morphological definition of that population. A first ring forms due to the lobes of FRII samples. A second ring forms due to the extended jet emission of FRI sources, which are some of the brightest regions of the given source image. This effect is likely

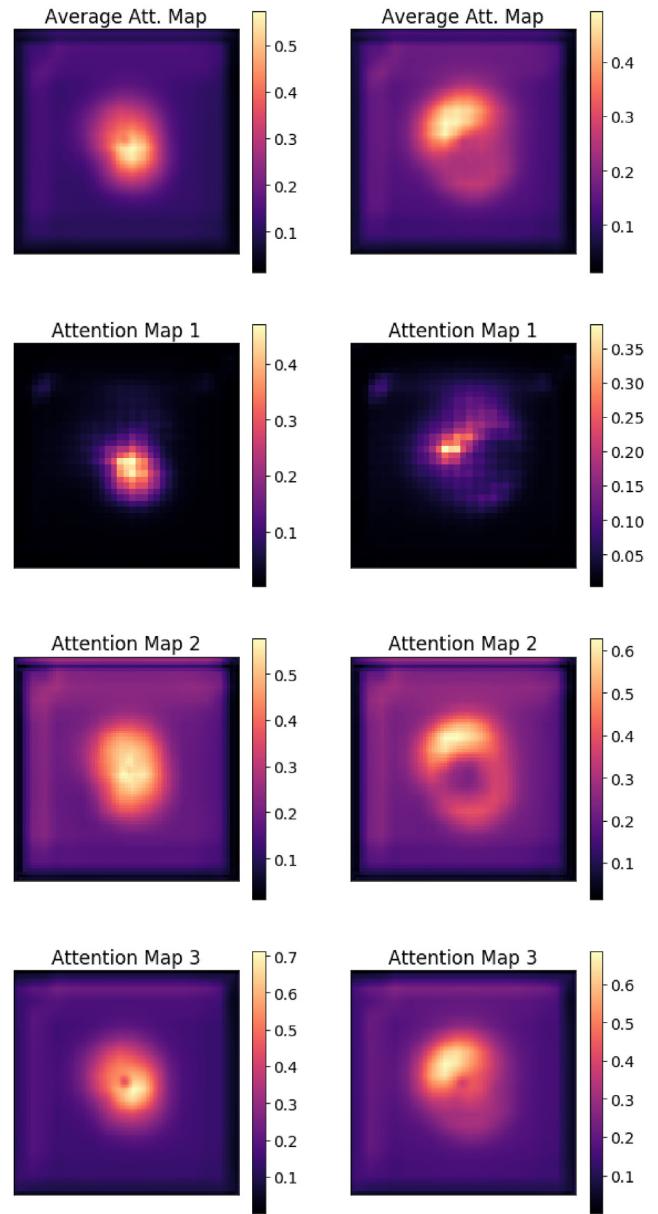


Figure 9. Mean pixel values of the average and individual attention maps of the test set after 100 fold augmentation. Left: FRI. Right: FRII.

enhanced due to the min-max normalization applied to each image. A third ring forms due to the peak emission of FRII sources often being the hot spots in the extended lobes, which tend to be much brighter than their jets.

Fig. 9 shows the difference in attention when separating the mean pixel intensities by class across the whole of the augmented MiraBest test set. The aggregate attention map shows a ring-like structure for FRII sources, which seems to stem primarily from attention gate 2. For the FRI sources, a centred Gaussian attention is more prevalent. As FRII sources are typically classified on their lobes and FRI sources are classified on the brightness of their central engines, this aligns well with how a human classifier would attend a data set on average.

By considering the aggregated attention across various subsets, a number of insights can be gained. Analogous to the discussion in previous section, Fig. 10 shows how the aggregate attention develops throughout training. At epoch 0 the aggregate attention is highly

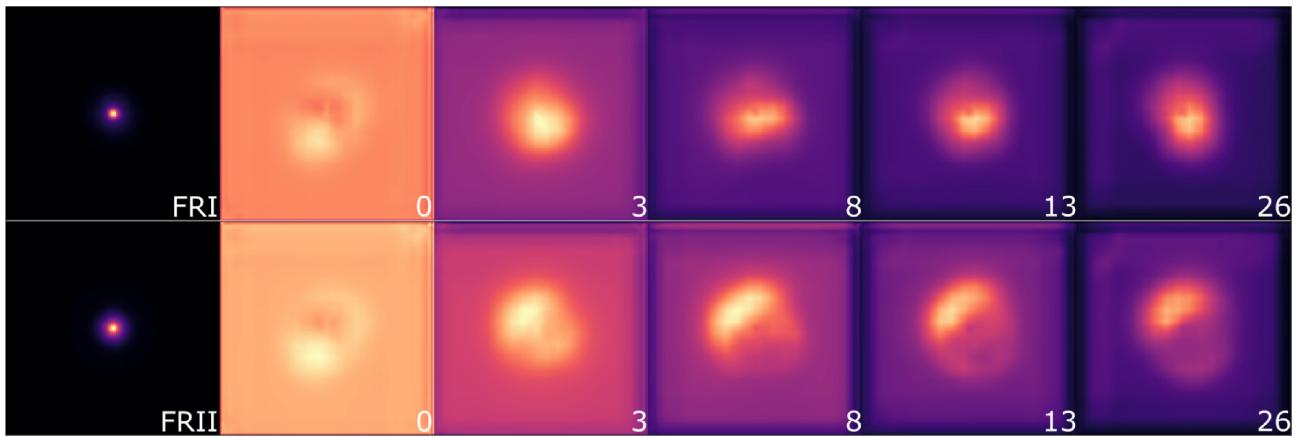


Figure 10. Mean pixel values of sources and attention maps throughout training at given epochs after 100-fold augmentation. See Fig. 7 for the equivalent plot with the example sources.

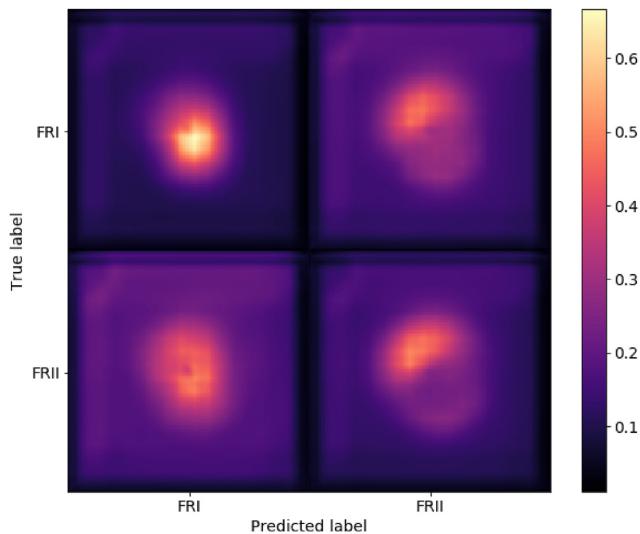


Figure 11. Distribution of attention maps according to the predicted and true labels of the respective sources.

similar across the two classes, but as the epochs progress the ring and Gaussian shapes develop in the aggregate attention maps of the respective classes, demonstrating how clearly the two classes are separated by the model’s attention before any fully connected layers are applied.

We also consider that it may be possible to see a difference in the aggregate attention maps for sources that are predicted correctly and those which the model predicts incorrectly. To visualize this, Fig. 11 shows a ‘confusion matrix’ of aggregated attention maps. It can be seen that the sources which were predicted incorrectly present less clearly defined attention with the largest difference being in the size of the central region and the direction of the offset of the bright-spot in the attended areas which aligns with the predicted class.

Similarly, the uncertainty in the data set becomes clear when considering the mean pixel intensities of the MiraBest-Certain and MiraBest-Uncertain attention maps and test data. Fig. 12 shows these maps and it can be seen that the attention maps are significantly less distinct for the MiraBest-Uncertain sources. The mean pixel intensities themselves can be separated by the peaks of the respective maps, which do not align with those of the confident FRI and FRII

sources. The FRII peak is expected to be higher, as FRII sources with brighter cores are expected to be ‘uncertain’. The uncertain FRI source mean intensity map has a higher peak than the certain FRI map. This may be due to the inherent classification bias, where fringe source cases with bright centres are more likely to be classified as FRI even though they are inherently uncertain.

Finally, the aggregate attention maps of the 16 models trained with each permutation of aggregation and normalization, introduced in Section 5.1, are shown in Fig. 13. In the case of the individual exemplar sources (see Fig. 5), the sigmoid models seem to be attending regions that an observer cannot clearly recognize; however, in the case of the aggregate attention maps, it is clear that they tend to highlight zero space around the source distributions, with little differences between the respective FRI and FRII distributions.

Similarly to the case of individual sources, the softmax normalized aggregate attention map is not helpful. Beyond the softmax models, all of the models presented demonstrate aggregate attention focused either on the central region of the image (range normalized models and some standardized models) or attend the zero spaces around the mean source pixel intensities (sigmoid models and some standardized models). Although knowing that the sigmoid models are attending regions representative of the mean source pixel intensities builds confidence in the models, their individual attention maps are not helpful for individual source analysis (see Fig. 5).

The value of these mean intensity and aggregated attention maps lies not only in the analysis of the model and understanding the difference in how it attends various subsets but also in the assistance they provide when developing new models, as they can be used to analyse the data set distributions and evaluate a model’s ability to generalize to unseen testing data.

7 INDIVIDUAL SOURCES

Using this work’s primary model, which implements range normalization, fine-tuned aggregation, and three attention gates trained on the MiraBest data set, we highlight some specific sources of potential interest. To do this, we evaluate the test set across 1000 random augmentations, as described in Section 4.3, and calculate the false classification rate for each test source.

By separating out sources that are classified incorrectly for more than 95 per cent of the augmentations, we highlight the objects which fundamentally do not align with the model’s learned classification.

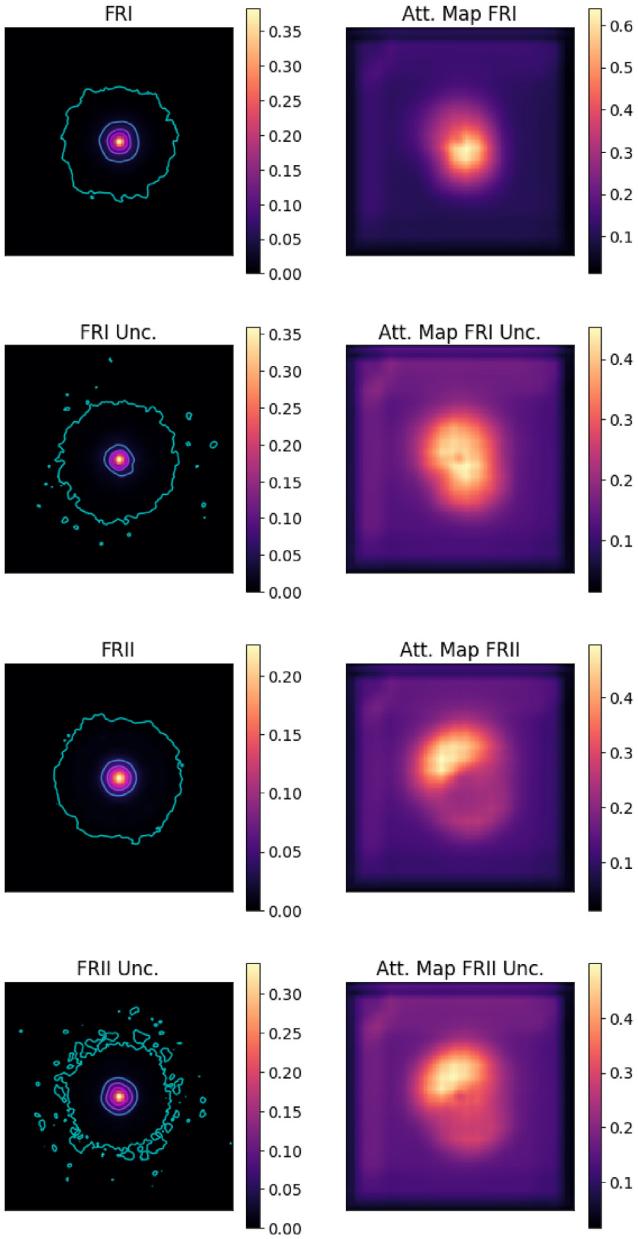


Figure 12. Mean source intensities and attention maps according to both class and certainty. With contours plotted alongside, the mean source mean intensity to highlight class differences at 0.001, 0.005, 0.01, 0.05, and 0.1, respectively.

Details of these sources are given in Table 10, and Fig. 14 presents their images and respective attention maps.

Sources A and B are the only certain FRI sources mis-classified as FRII sources at a higher rate than 95 per cent by the model. The attention map of source A shows that the model is attending the features of the bright source near the centre of the image; however, it is unclear whether this bright source is aligned with the host galaxy for this source. If it is, then this source seems to be mis-labelled, as the only visible jet would be edge brightened. If the bright source itself is the central engine, then the source would be classified as an FRI. The model's training procedure could be adapted to allow for significantly off-centre sources by augmenting the data such that the central engine of each source could be anywhere on the image. This

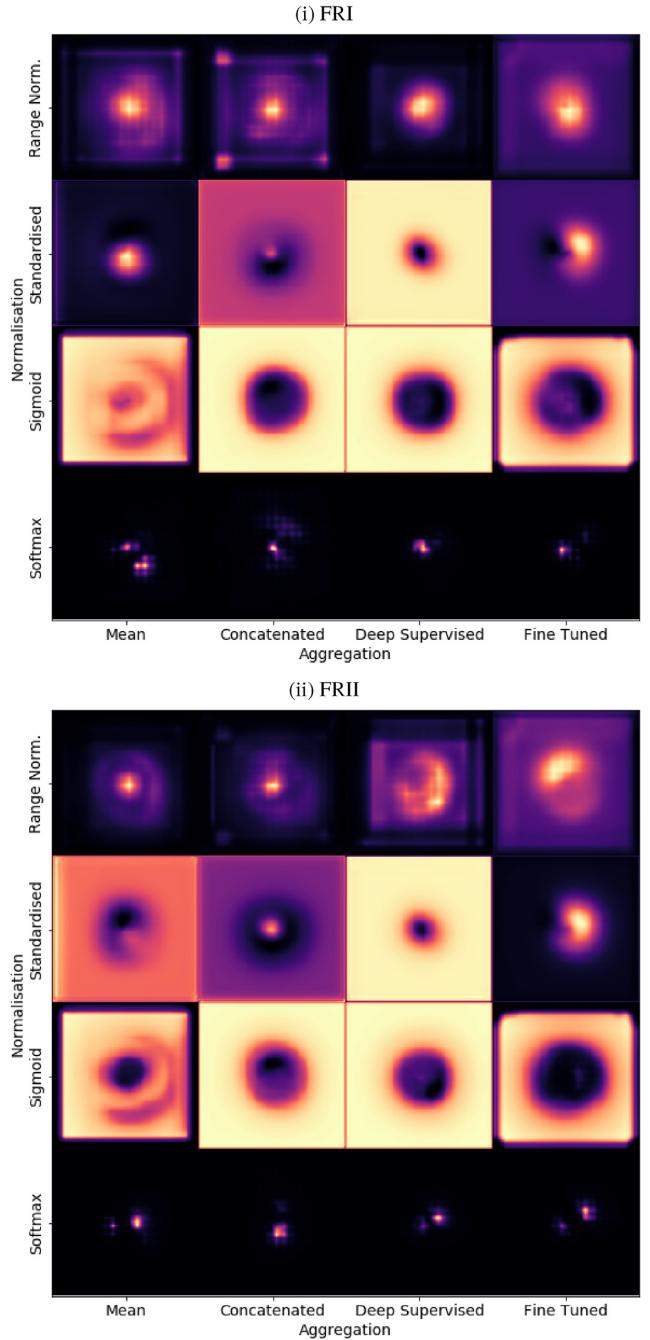


Figure 13. Aggregated attention maps of models implemented with various aggregations and normalizations on the (i) FRI and (ii) FRII MiraBest data sets.

is an example of how attention maps can help inform the training of the model and the selection of the training data.

The class of source B is not immediately clear from a simple visual inspection. The jets do not have individual bright spots, but rather have bright regions which makes the classification less clear. It may be that the training data simply did not contain any cases of FRI sources where the jets showed bright extended regions (instead of hot spots) which extended into the largest part of the extent of the source. This could be remedied by training with more data, which would hopefully 'fill in the gaps' which the model still has.

Table 10. Details for the sources with a mis-classification rate above 95 per cent, as depicted in Fig. 14. Separated by dashes, the SDSS IDs are composed of the source’s plate ID, Julian date and Fibre ID.

Source	SDSS ID	RA	Dec.	Redshift	Extent (arcsec)	MiraBest Class
A	1737-53055-197	07 ^h 48 ^m 18 ^s .9	+45°44'46"	0.1850	228.03	Certain FRI
B	0875-52354-521	10 ^h 40 ^m 22 ^s .5	+50°56'25"	0.1539	44.61	Certain FRI
C	1618-53116-159	11 ^h 29 ^m 54 ^s .5	+06°53'12"	0.1162	120.67	Uncertain FRI
D	2598-54232-336	12 ^h 24 ^m 46 ^s .7	+18°25'32"	0.1690	14.13	Uncertain FRI
E	2239-53726-557	12 ^h 51 ^m 57 ^s .1	+30°09'26"	0.2235	11.94	Uncertain FRI
F	2510-53877-594	11 ^h 50 ^m 03 ^s .7	+25°39'26"	0.1561	128.69	Uncertain FRI
G	1690-53475-090	16 ^h 49 ^m 24 ^s .0	+26°35'03"	0.0545	68.49	Uncertain FRI
H	2203-53915-518	16 ^h 30 ^m 16 ^s .6	+14°35'11"	0.2790	106.72	Uncertain FRII

Sources C–G are classed as uncertain FRI sources. Sources D, E, and F are not well resolved and it is difficult to specify which class they belong to (assuming more distant flux in the image is not from the central radio galaxy itself). For source C, the model may in fact be classifying a single lobe of the radio galaxy, as indicated by the enhancement in the attention map to the top left of the image. The model is not aware that the source has been extracted to be at the centre of the image and may thus classify the lobe as an FRII radio galaxy.

Source G is not a bent source, nor is it unresolved or strange in any other way. It does however seem difficult (on first viewing) to see whether or not it is an FRI or FRII, as the distance ratio is not clear on first viewing, but would require explicit measurement.

Source H is labelled as an uncertain FRII within the MiraBest data set, and is misclassified with a rate of over 95 per cent as an FRI by the model. The model tends to overclassify sources as FRII sources due to the slight imbalance in the training data, which makes this object a notable exception as the only FRII source mis-classified as FRI. A central bright-spot is clearly visible, but unlike the exemplar FRII source displayed in Fig. 4, source H shows no clear bright-spots in its wide-angled lobes. Mis-classification in this case may arise from the unusually bright central source, which itself may be a chance alignment with another galaxy along the line of sight. Alternatively, it may be due to confusion with the secondary source at the bottom right of the image, which is associated with an enhanced area of attention.

Some of these sources may require more intentional efforts to be well understood and we note that outlining what could have caused the model to classify these sources incorrectly is not an attempt at physical analysis. This evaluation of the outlying data samples is meant to serve in two aspects. First, it is meant to highlight sources which may be of interest for improving the performance of the model. Secondly, this evaluation is meant to demonstrate the value of attention and machine learning for extracting individually abnormal sources, even from a seemingly simple data set, in the same way that Sasmal et al. (2020) manually selected and highlighted abnormal sources from the LOTSS data release 1. Future applications of this technique might also consider including data sets at other wavebands to clarify these aspects, in the same way that Wu et al. (2019) combined FIRST and WISE data for the CLARAN classifier.

8 CONCLUSIONS

In this work, we introduce attention as a state of the art mechanism for classification of radio galaxies using convolutional neural networks. We present an attention-based model that performs on par with previous classifiers while using over 50 per cent fewer parameters than the next smallest classic CNN application in this field. Furthermore, the AG-CNN presented in this work provides the additional benefit

of visualizing the salient regions used by the model in each case to make individual classifications.

The primary model in this work was implemented using range normalization, fine-tuned aggregation and three attention gates. The model primarily attends the central engine for FRI sources, and increasingly attends the outer regions (i. e. lobes) of FRII sources. We observe that the salient regions identified by the attention gated model align well with the regions an expert human classifier would attend to make equivalent classifications. This includes both the central engine of the respective sources, the hot spots and the lobes of the source itself. This is also shown to be a learned trait of the model, by explicitly considering how the model’s attention develops throughout training.

We also investigate how the selection of normalization and aggregation methods used in attention-gating affect the output of individual models, using both quantitative evaluation metrics and the resulting attention maps to determine how employable each resulting model may be. Although the selection of such parameters minimally affects the model’s performance, it can adversely affect individual models with regard to the interpretability of their respective attention maps. By selecting a model which aligns with how astronomers classify radio sources, the user can then employ the model to investigate what features the model is using to make classifications, and thus investigate how future models may be improved. We find that the softmax normalization and concatenated aggregation methods provide the best model performance, but suggest that the range normalization and fine-tuned aggregation methods provide the user with significantly improved attention maps at the cost of a minimal difference in performance. Similarly we find that the inclusion of a third attention gate does not contribute significantly to model performance but does aid in the interpretability of the resulting attention maps.

We evaluate the average performance of models across the entire data set and use the example of individual sources in order to illustrate how attention maps can help the user engage in classification. By considering the aggregate attention maps across the augmented test set we note significant differences between various subsets of the data including the fundamental class division between FRI and FRII as well as that between correctly and incorrectly classified sources within each class.

Finally, we present a method through which deep learning models can highlight individual sources for further study by extracting test sources that were found to be significantly misaligned with the predictions of the trained model. In these cases, the availability of the attention maps can be used to examine the cause of the misclassification in each case, as well as to understand of how complex the data might be, and how difficult even binary classification can become in certain cases. When applied to labelled data, this approach might also be used to select potential targets for future

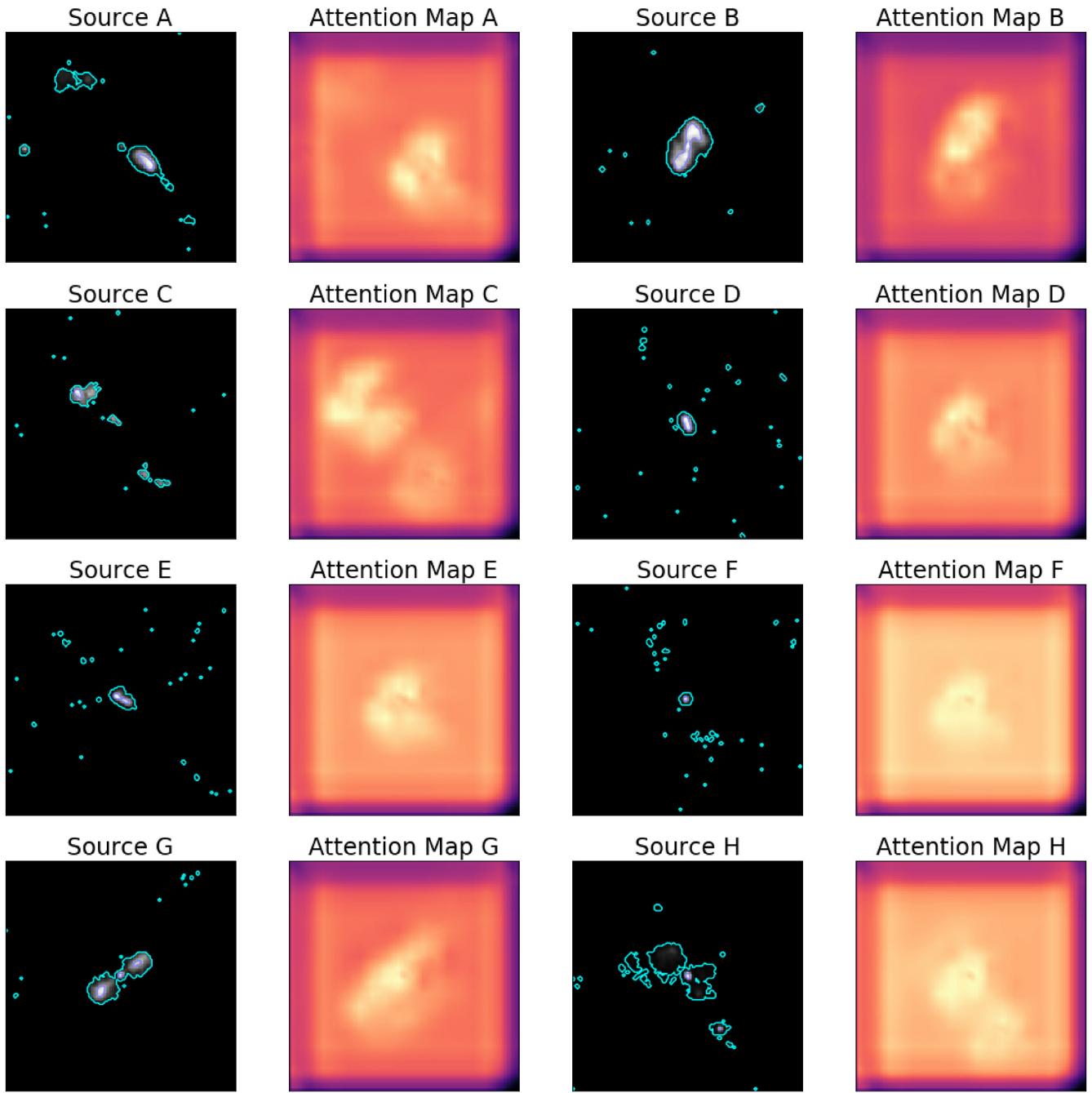


Figure 14. Sources that have a false classification rate above 95 per cent. The pre-processed source image, with a contour highlighting the sigma clipped zero space, and their respective attention maps. Sources are grouped by class label and certainty. Details for each source can be found in Table 10.

research by identifying sources that the model seems to consider as significant outliers.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewer for useful feedback that improved the contents of this paper. MB gratefully acknowledges support from the University of Manchester STFC CDT in Data Intensive Science, grant number ST/P006795/1. AMS gratefully acknowledges support from an Alan Turing Institute AI Fellowship EP/V030302/1. FP gratefully acknowledges support from STFC and IBM through the iCASE studentship ST/P006795/1.

DJB gratefully acknowledges support from STFC and the Newton Fund through the DARA Big Data program under grant number ST/R001898/1.

DATA AVAILABILITY

All code and trained models from this work are publicly available on Github at the following address: github.com/mb010/AstroAttention. MiraBest is available on Zenodo: 10.5281/zenodo.4288837. FR-DEEP is available on Zenodo: 10.5281/zenodo.4255826.

REFERENCES

- Aguado D. S. et al., 2019, *ApJS*, 240, 23
- Aniyan A. K., Thorat K., 2017, *ApJS*, 230, 20
- Bahdanau D., Cho K., Bengio Y., 2015, CoRR, abs/1409.0473
- Bai Y., Liu J. F., Wang S., 2018, *Res. Astron. Astrophys.*, 18, 118
- Ba J., Mnih V., Kavukcuoglu K., 2014, CoRR, abs/1412.7755
- Baldi R. D., Capetti A., Giovannini G., 2015, *A&A*, 576, A38
- Baldi R. D., Capetti A., Giovannini G., 2016, *Astron. Nachr.*, 337, 114
- Baldi R. D., Capetti A., Giovannini G., 2019, *MNRAS*, 482, 2294
- Beardsley A. P. et al., 2019, *Publ. Astron. Soc. Aust.*, 36, e050
- Becker R. H., White R. L., Helfand D. J., 1995, *ApJ*, 450, 559
- Best P. N., Ker L. M., Simpson C., Rigby E. E., Sabater J., 2014, *MNRAS*, 445, 955
- Bonafede A., Feretti L., Murgia M., Govoni F., Giovannini G., Dallacasa D., Dolag K., Taylor G. B., 2010, *A&A*, 513, A30
- Capetti A., Massaro F., Baldi R. D., 2017, *A&A*, 598, A81
- Capetti A., Massaro F., Baldi R. D., 2020, *A&A*, 633, A161
- Cervantes J., Garcia-Lamont F., Rodriguez-Mazahua L., Lopez A., 2020, *Neurocomputing*, 408, 189
- Chen L., Zhang H., Xiao J., Nie L., Shao J., Liu W., Chua T. S., 2017, *Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017*. IEEE, Honolulu, HI, p. 6298
- Clarke A. O., Scaife A. M. M., Greenhalgh R., Griguta V., 2020, *A&A*, 639, A84
- Das P., Sanders J. L., 2019, *MNRAS*, 484, 294
- Deng J., Dong W., Socher R., Li L., Li K., Li F.-F., 2009, 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Miami, FL, p. 248
- Fanaroff B. L., Riley J. M., 1974, *MNRAS*, 167, 31P
- Gaia Collaboration et al., 2016, *A&A*, 595, A1
- Galassi A., Lippi M., Torroni P., 2019, *IEEE transactions on neural networks and learning systems*
- Gendre M. A., Best P. N., Wall J. V., 2010, *MNRAS*, 404, 1719
- Gendre M. A., Wall J. V., 2008, *MNRAS*, 390, 819
- Godfrey L. E. H., Morganti R., Brienza M., 2017, *MNRAS*, 471, 891
- Gopal-Krishna, Wiita P. J., 2000, *A&A*, 363, 507
- Govoni F. et al., 2010, *A&A*, 522, A105
- Gron A., 2017, Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 1st edn. O'Reilly Media, Inc, Sebastopol, CA
- Hardcastle M. J., 2018, *MNRAS*, 475, 2768
- Hardcastle M. J., Croston J. H., 2020, *New Astron. Rev.*, 88, 101539
- Hartley P., Flamary R., Jackson N., Tagore A. S., Metcalf R. B., 2017, *MNRAS*, 471, 3378
- Ineson J., Croston J. H., Hardcastle M. J., Kraft R. P., Evans D. A., Jarvis M., 2015, *MNRAS*, 453, 2682
- Itti L., Koch C., 2001, *Nature Rev. Neurosci.*, 2, 194
- Jetley S., Lord N. A., Lee N., 2018, CoRR, abs/1804.02391
- Johnston S. et al., 2008, *Exp. Astron.*, 22, 151
- Kapińska A. D. et al., 2017, *ApJ*, 154, 253
- Lahav O., Naim A., Sodré L., Storrie-Lombardi M. C., 1996, *MNRAS*, 283, 207
- Lindsay G. W., 2020, *Frontiers Comput. Neurosci.*, 14, 29
- Louppe G., 2014, preprint (arXiv:1407.7502)
- Lukic V., Brüggen M., Mingo B., Croston J. H., Kasieczka G., Best P. N., 2019, *MNRAS*, 487, 1729
- Mahatma V. H. et al., 2019, *A&A*, 622, A13
- Ma Z. et al., 2019, *ApJS*, 240, 34
- Miller A. S., 1993, *Vistas in Astronomy*, 36, 141
- Mingo B. et al., 2019, *MNRAS*, 488, 2701
- Miraghaei H., Best P. N., 2017, *MNRAS*, 466, 4346
- Netzer H., 2015, *ARA&A*, 53, 365,
- O'Sullivan S. P. et al., 2019, *A&A*, 622, A16
- Porter F., 2020, MiraBest Batched Dataset, 10.5281/zenodo.4288837
- Qu M., Shih F. Y., Jing J., Wang H., 2006, *Sol. Phys.*, 237, 419
- Raccanelli A. et al., 2012, *MNRAS*, 424, 801
- Sasmal T. K., Bera S., Pal S., Mondal S., 2020, *J. Phys. Conf. Ser.*, 1579, 012021
- Schlemper J., Oktay O., Chen L., Matthew J., Knight C. L., Kainz B., Glocker B., Rueckert D., 2018, MIDL 2018, openreview.net, conference track, non-archival (arXiv:1804.05338)
- Schlemper J., Oktay O., Schaap M., Heinrich M., Kainz B., Glocker B., Rueckert D., 2019, *Med. Image Anal.*, 53, 197
- Schoenmakers A. P., De Bruyn A. G., Röttgering H. J., Van Der Laan H., Kaiser C. R., 2000, *MNRAS*, 315, 371
- Seymour N. et al., 2020, *Publ. Astron. Soc. Aust.*, 37, e013
- Shimwell T. W. et al., 2019, *A&A*, 622, A1
- Simonyan K., Zisserman A., 2014, Very Deep Convolutional Networks for Large-Scale Image Recognition, CoRR, abs/1409.1556
- Smithand M. D., Donohoe J., 2019, *MNRAS*, 490, 1363
- Stollenga M. F., Masci J., Gomez F. J., Schmidhuber J., 2014, CoRR, abs/1407.3068
- Tang H., Scaife A. M. M., Leahy J. P., 2019, *MNRAS*, 488, 3358
- Taylor A. R., Jarvis M., 2017, IOP Conf. Ser.: Mater. Sci. Eng., Vol. 198. Réunion Island, France
- Torresi E., Grandi P., Capetti A., Baldi R. D., Giovannini G., 2018, *MNRAS*, 476, 5535
- Van Haarlem M. P. et al., 2013, *A&A*, 556, A2
- Wu C. et al., 2019, *MNRAS*, 482, 1211
- Xu K., Ba J. L., Kiros R., Cho K., Courville A., Salakhutdinov R., Zemel R. S., Bengio Y., 2015, 32nd International Conference on Machine Learning, ICML 2015. Lille, France, JMLR.org, p. 2048
- Zhou H., Desimone R., 2011, *Neuron*, 70, 1205

APPENDIX: CLASSIFICATION EVALUATION METRICS

Classifications made by any model will be true (T) or false (F) as well as positive (P) or negative (N) for a given class. Knowing this, models can be evaluated by how many predictions fall into each of the four subsets: TP, TN, FP, and FN. Gron (2017) contains a helpful overview of the metrics introduced here.

A.1 Accuracy

Accuracy $\in [0, 1]$ is the ratio between correct predictions and all predictions. For data sets where the class sizes are not equal, accuracy should be calculated on a per class basis:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \quad (\text{A1})$$

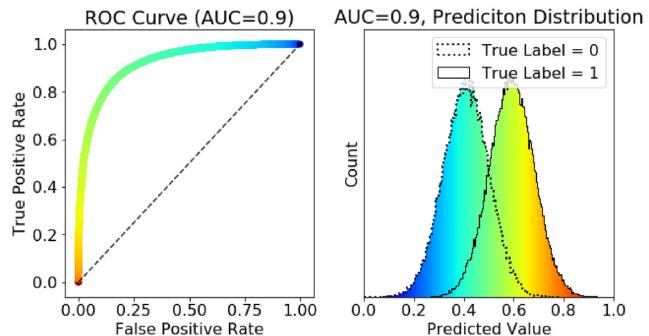


Figure A1. ROC curve and the distributions which created it. Each distribution is a normal distribution of predicted class values in [0,1]. The threshold of a given point on the ROC curve is equivalent to the vertical colour of the distributions. The largest class separation occurs at PredictedValue ≈ 0.5 (green), which provides the optimal trade-off between TPR and FRP in this example case.

A.2 Precision

Precision $\in [0, 1]$ is the ratio of positive classifications of all the positive classifications made:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \quad (\text{A2})$$

A.3 Recall

Recall $\in [0, 1]$ is the proportion of positive samples which are classified positively. Recall is equivalent to class specific accuracies in the binary case:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (\text{A3})$$

A.4 F1 Score

$F_1 \in [0, 1]$ is the harmonic mean of precision and recall, averages their respective reciprocals. This is done to ensure that if either precision or recall is low, the F1 score suffers:

$$F_1 = \frac{2}{\text{Precision}^{-1} + \text{Recall}^{-1}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}. \quad (\text{A4})$$

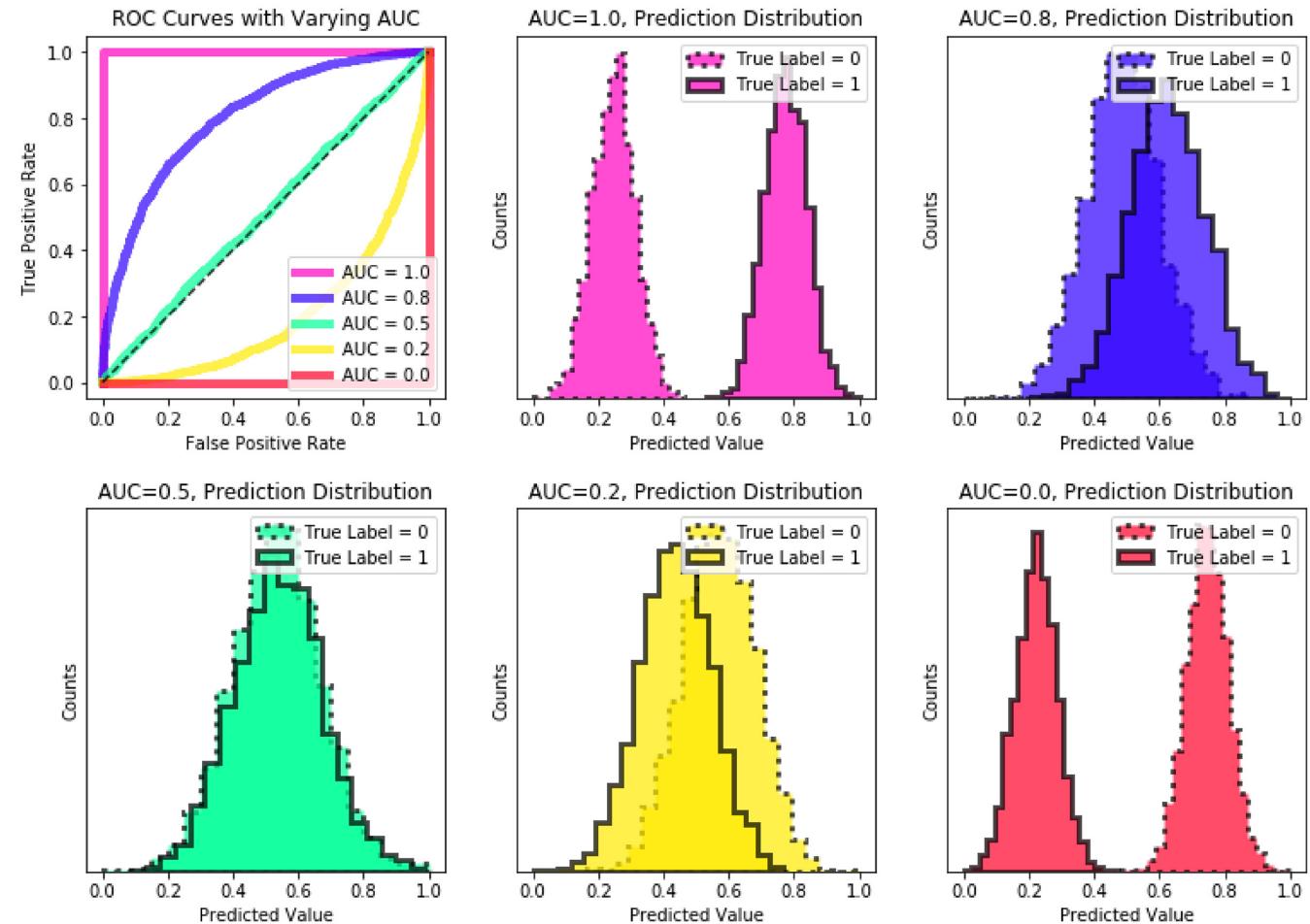


Figure A2. Visualizing how AUC relates to the predictions of a given model. Each distribution relates to one ROC curve, and the AUC score is listed with each of the distributions and curves, respectively.

This paper has been typeset from a TeX/LaTeX file prepared by the author.

A.5 ROC

The Receiver Operator Characteristic (ROC) curve is the name given to the curve which results from considering true positive rates (equivalent to recall) and the false positive rates:

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}}, \quad (\text{A5})$$

at various thresholds. Thresholds are the value at which the two classes are separated. This is often 0 in the case of predictions in the range of $[-1, 1]$, or 0.5 in the case of $[0, 1]$. The threshold values for a example prediction distribution are visualized in Fig. A1. Given a binary classification, the ROC curve plots the recall of one class against the recall of the other.

A.6 AUC

Area Under Curve, or AUC $\in [0, 1]$, is the area under the ROC curve. It is a measure of how well the models predictions have separated the two classes. Examples of AUC values, along with the respective distributions and ROC curves, are visualized in Fig. A2. Any value below 0.5 generally would indicate some implementation error, as the model seemingly separates the classes, but assigns the labels in reverse.