Uncertainty calibration for group-equivariant Bayesian CNNs in radio galaxy classification

A. Millicheap (10323486) and S. Hudson

Supervised by: A. M. M. Scaife

Abstract

In this work the Monte Carlo dropout method is used to quantify the degree of uncertainty in a LeNet-5 Convolutional Neural Network used in automated machine learning predictions of Fanaroff-Riley galaxy classifications. The method is used to compare this model to equivalent models of varying group-equivariance. We consider the model performance and uncertainty calibration for these models using different uncertainty and calibration metrics. We find that an overlap index is the best calibrated uncertainty metric and best represents the predicted error rate of the network for the *MiraBest** data set. We demonstrate that the implementation of group-equivariance to the neural network increases the accuracy of models, however there is no consistent trend found in their uncertainty calibration. We consider the effectiveness of implementing spectral normalisation as a method to improve model robustness. It is found that there is no consistent or singificant improvement to model robustness from implementation of spectral normalisation. By retraining the network multiple times it is demonstrated that there is an uncertainty in the UCE calibration metric of $\sim 17\%$ using the overlap index as an uncertainty metric and therefore it is likely that any trends in the uncertainty calibration would be masked by this random variation.

1. Introduction

As neural networks (NNs) become more integrated into critical decision-making applications, it is essential to test the robustness of each model in order to evaluate the efficacy of new models and optimise existing ones. It is known that NNs frequently make poor predictions with a high predictive probability, creating a disparity between the model confidence and the true error in a decision[19]. It is necessary that the influence of each datapoint in the model training can be weighted accordingly so that poor-quality and out-of-distribution data can be discounted. It is therefore important to evaluate the confidence calibration of the model, assessing the ability of the model to distinguish accurate predictions from inaccurate ones.

The field of research of radio astronomy has seen drastic changes in recent years. The beginning of the development of the Square Kilometer Array (SKA) and the collection of vast quantities of data collected by new facilities such as the e-MERLIN array[16], the Low-Frequency Array(LOFAR) [44], the MeerKAT telescope [21] and the Australian SKA Pathfinder telescope (ASKAP)[24] have led to a revolution due to an increase in the ability to collect and store data. For example, the Australian Square Kilometre Array Pathfinder is expected to catalogue approximately 70 million galaxies [1]. However, this increase in data has required a reevaluation of the approach to galaxy cataloguing from the traditional *'by eye'* method. The introduction of autonomous cataloguing through Machine Learning (ML) methods has allowed radio astronomers to capitalise on these changes and accelerate progress.

Fanaroff and Riley introduced a classification system to distinguish active galaxy nucleus (AGN) powered radio galaxies based on the ratio of the distance between the points of greatest brightness on either side of the nucleus and the total extent of the source[12]. Fanaroff-Riley Class I (FR-I) galaxies

are classified by a ratio of less than 0.5 and typically feature diffuse jets. Fanaroff-Riley Class II (FR-II) galaxies are classified by a ratio of greater than 0.5 and typically feature bright "hot spots" at a distance from the nucleus.

Convolutional neural networks (CNNs) have proven extremely well-suited to morphological object classification due to their translational equivariance which allows them to encode information about the position of an object in an image[40][2]. The structure of a CNN is made up of a combination of convolutional layers, activation functions and pooling layers in a feedforward network with multiple layers[37] . Using a backpropagation training process, CNNs optimise the weights of a convolutional kernel in order to best match known classifications. The use of CNNs was first introduced to FR galaxy classification by Aniyan & Thorat[1]. A number of notable works followed using various deep learning methods, for example the work of Lukic et al.[32], Wu et al.[7] and Tang et al.[42] Recent works by Bowles et al. found that radio galaxies could be classified with similar accuracy to previous methods with $\sim 50\%$ fewer learning parameters by using an attention-gated CNN[5]. While a standard CNN is translationally equivariant, one limitation is a lack of equivariance to rotational and reflectional transformations.

Equivariance under given transformations is essential for image classification in order for the network to generalise to a test data set. Numerous CNN models have been developed to implement equivariance to different symmetry groups. A network is equivariant to a transformation if, under the transformation of the input data, the corresponding representation transforms in a linear manner. The simplest way that approximate equivariance can be achieved is through data augmentation[41]. Random rotational transformations are applied to the image in order to encourage the network to correctly classify the training data regardless of orientation. Provided the network has sufficient capacity it should learn approximate equivariance, however this does not always generalise to a test data set[45]. Dieleman et al. introduced four additional layers to their CNN to integrate discrete rotational equivariance by concatenating the outputs of convolutional layers for duplicates of a single data sample at various orientations[11]. However, this approach requires a replicated model weight for each discrete rotation, making it computationally expensive.

A more efficient method introduced by Cohen & Welling preserves equivariance to a particular group of transformations using group convolution layers[8]. These Group equivariant Convolutional Neural Networks (G-CNNs) extend upon the translational equivariance of standard CNNs which is ensured through weight sharing. By exploiting symmetries, G-CNNs produce steerable representations and therefore allow filters to be applied in alternative orientations. The network therefore has a lower computational cost as it benefits from increased parameter sharing.

The E(2) Euclidian group is the group of isometries describing translations, reflections and rotations in the $\mathbb{R}^2$ plane. E(2) equivariance is important in automated galaxy classification for the NN to generalise to real-world data as galaxies do not have a specified position or orientation in 3D space. Scaife & Porter used a G-steerable CNN to conserve E(2) isometries in FR galaxy classification[40]. The work applied the Monte Carlo (MC) dropout method in accord with Krizhevsky et al.[28] as a Bayesian approximation to quantify model uncertainty, in order to assess the rotational and reflectional invariance of various models with different levels of equivariance. Further work by Mohan et al. used variational inference in place of the MC dropout to implement a fully Bayesian CNN[35]. This was used to compare model uncertainty for test data sets with different human classifications applied.

In this work the MC dropout method is used as a Bayesian approximation to calculate the uncertainty of G-steerable CNNs with varying orders of equivariance. The method consists of randomly omitting 50% of the nodes for an NN at each layer over N stochastic forward passes on a test sample set. The mean class prediction provides a softmax probability reflecting the confidence in the result. The uncertainty calibration of these networks is quantified and evaluated for various uncertainty metrics compared against the error rate of the model. The results are visualised using reliability diagrams, as displayed in Figures 4,

5 and 6. This allows us to evaluate whether the confidence values for each network are representative of true probabilities and provides insight into the factors which cause their miscalibration. Finally, spectral normalisation is applied in an attempt to recalibrate the models and improve their robustness.

2. Uncertainty Quantification in Deep Learning

Combining Bayesian inference with neural networks has produced a novel form of NN known as Bayesian Neural Networks (BNNs). Standard NNs optimise model parameters by maximising the likelihood for a data set x and a model parameterised by weights W, $\mathcal{L}(x|W)$ to produce point-wise weights and outputs. Alternatively, BNNs deduce a posterior distribution $P(W|x)$ using Bayes' rule:

$$P(W|x) \propto \mathcal{L}(x|W) \times \pi(W), \tag{1}$$

where $\pi(W)$ is the prior distribution of the model for a given set of weights. By assuming a prior distribution and a form for the posterior distribution, a posterior conditioned on training data can be used to calculate a predictive posterior distribution for a class prediction y on new data, x^*:

$$P(y|x^*, x) = \int P(y|x^*, W)P(W|x)dW. \tag{2}$$

Bayesian methods have proven useful in NNs[43][22] due to their application to real-world parameter uncertainties, as well as providing improved accuracy and robustness for smaller datasets [14]. However, while the benefit of BNNs is considerable, numerous techniques have been introduced to reduce their high computational complexity. Recent BNN research has focused on variational inference (VI) [18] and Markov Chain Monte Carlo (MCMC) methods [6]. Jia et al. introduced feature decomposition and memorisation features, which was successful in reducing energy consumption by 73%[23].

Gal and Ghahramani recently provided mathematical proof for how the MC dropout method can be used to approximate a Bayesian predictive posterior distribution[15]. Equation (2) can be approximated as a sum:

$$P(y|x^*, x) = \frac{1}{N} \sum_{i=1}^{N} P(y|x^*, w^{(i)}), \tag{3}$$

where $w^{(i)} \in W$. The MC dropout method assumes a Gaussian prior distribution, $\mathcal{N}(0, 1)$, and assumes the form of the posterior distribution to be a Gaussian Mixture Model (GMM). Equation (3) can therefore be adjusted to display the underlying distribution for MC dropout:

$$P(y|x^*, x) = \frac{1}{N} \sum_{i=1}^{N} z_1\mathcal{N}(\nu, \epsilon_1) + z_2\mathcal{N}(\nu, \epsilon_2), \tag{4}$$

where z_i are Bernoulli random variables with a 50% chance of being 0 or 1, ν is the calculated mean of the distribution and ϵ_i is the covariance matrix of the *i*th Normal distribution.

Sources of uncertainty in deep learning models can be classified as either aleatoric or epistemic [20]. Aleatoric uncertainties (AU) refer to the variation in results due to irreducible statistical factors. Epistemic uncertainties (EU) are derived from a lack of knowledge and are reducible provided additional information can be obtained. The sum of these uncertainties gives the predictive uncertainty (PU):

$$PU = EU + AU. \tag{5}$$

These uncertainties can be quantified by the metrics outlined in the remainder of this section.

2.1. Predictive Entropy

Predictive entropy describes the amount of information contained within the predictive distribution. It provides a measure of the total predictive uncertainty of a model:

$$\mathbb{H}(y|x^*, x) = -\sum_c P(y = c|x^*, x)\log P(y = c|x^*, x), \tag{6}$$

for all classes c which can be taken. The metric's value has a minimum of zero and maximum value of $\frac{1}{\ln 2}$. Using MC samples, the probability distributions are derived from averaging over N stochastic passes and therefore we approximate Equation 6 as:

$$\mathbb{H}(y|x^*, x) = -\sum_c \left(\frac{1}{N}\sum_{i=1}^{N} P(y = c|x^*, w^{(i)})\right) \log\left(\frac{1}{N}\sum_{i=1}^{N} P(y = c|x^*, w^{(i)})\right), \tag{7}$$

2.2. Mutual Information

Mutual information provides a metric for the epistemic uncertainty in a given model. Test data points that give a maximal value for mutual information are uncertain on average but have many results with low uncertainty values. For a set of model weights, W, it is given by the difference between the entropy over the training data set and the expectation value of the entropy over W:

$$\mathbb{I}(y|x^*, x) := \mathbb{H}(y|x^*, x) - \mathbb{E}(\mathbb{H}(y|x^*, x)), \tag{8}$$

which can be approximated for MC dropout according to the method of Gal[13] as:

$$\begin{aligned}\mathbb{I}(y, w|x^*, x) = -\sum_c \left(\frac{1}{N}\sum_{i=1}^{N} P(y = c|x^*, w^{(i)}\right) \log\left(\frac{1}{N}\sum_{i=1}^{N} P(y = c|x^*, w^{(i)})\right) \\ +\frac{1}{N}\sum_{c,N} P(y = c|x^*, w^{(i)})\log P(y = c|x^*, w^{(i)})\end{aligned} \tag{9}$$

2.3. Average Entropy

The average entropy can be used as a metric to analyse the aleatoric uncertainty of a neural network. Given by the expected value of predictive entropy, this can be extended to the case of MC dropout:

$$\mathbb{E}(\mathbb{H}(y|x^*, x)) = -\frac{1}{N}\sum_{c,N} P(y = c|x^*, w^{(i)})\log P(y = c|x^*, w^{(i)}) \tag{10}$$

2.4. Overlap Index

Pastore & Calcagne[39] introduced a distribution-free overlap index as a quantifiable metric for the similarity between samples. The overlap index offers an alternative to the maximum softmax output as a metric for the degree of certainty in a classification. The index is based upon the separation between the posterior distribution and the alternative classes. If the distribution is well-separated from alternative classes then the overlap is small and the uncertainty is low. If the distribution overlaps considerably with other target classes then the overlap index is large and this indicates a high uncertainty. Figure 1 displays

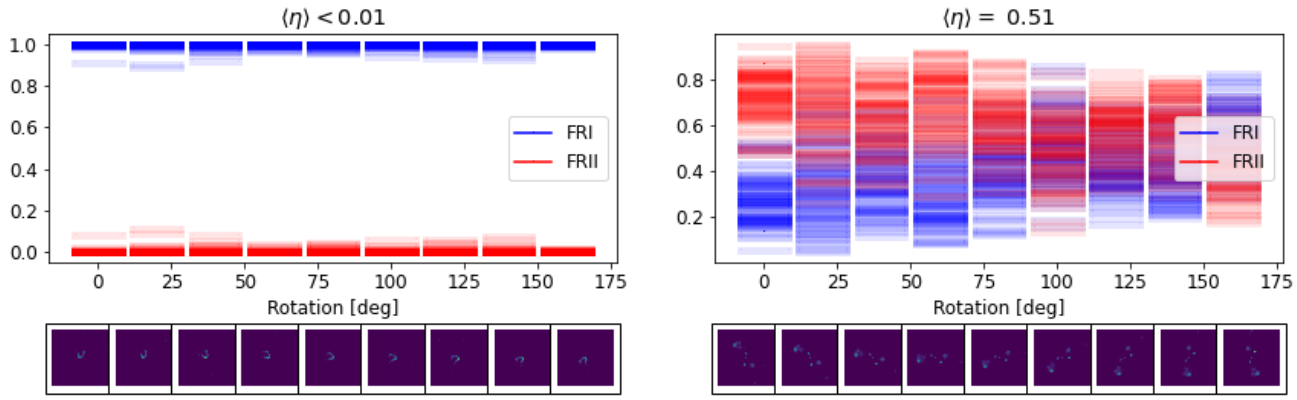

Figure 1: Graphical representation of the spread of Softmax outputs for D_{16} model outputs for two example samples over 100 forward passes as a function of rotation angle as the image is rotated 180 degrees in 20 degree increments. The average overlap is indicated above each graph. The mean overlap, $\langle\eta\rangle$ is displayed above either graph. Images of the galaxy at each orientation is displayed underneath. The sample with average overlap $\eta < 0.01$ (left) has well-separated classes and the model is therefore relatively confident. On the other hand, the classification for the sample with average overlap $\eta = 0.51$ (right) is not well-defined and therefore has a much higher uncertainty.

how an overlap index can represent model uncertainty. The overlap is calculated by first calculating the local density at a location z for each class based on a Gaussian kernel density estimator defined by:

$$f_x(z) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\beta\sqrt{2\pi}}e^{\frac{-(z-x_i)^2}{2\beta^2}} \tag{11}$$

$$f_y(z) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\beta\sqrt{2\pi}}e^{\frac{-(z-y_i)^2}{2\beta^2}} \tag{12}$$

where $\beta = 0.1$. The overlap index, η, is then calculated from:

$$\eta = \sum_{i=1}^{N_z}\min[f_x(z_i), f_y(z_i)]\delta z \tag{13}$$

where $\{z_i\}_{i=1}^{N_z}$ defines N_z evenly spaced steps in the range [0,1] and $\delta z = z_i$.

3. Uncertainty Calibration in Deep Learning

When evaluating the efficacy of a model, it is important to consider how well-calibrated the model's predicted uncertainty is to the actual error rate in order to credibly evaluate the confidence in predictions when applied to future unclassified data. A calibration metric allows the evaluation of statistical consistency between predictions and data by quantifying the correlation between a model's confidence and the accuracy of it's predictions. Reliability diagrams were introduced by DeGroot & Fienberg [10] as a visual representation for model calibration by plotting test accuracy as a function of model confidence. Values are plotted using a binned histogram made up of M bins containing $\frac{n}{M}$ samples, where n is the total number of samples. For a perfectly calibrated model the identity function will be plotted. The confidence value for a bin is calculated from the mean confidence value of the samples contained within

it:

$$conf(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i \tag{14}$$

where $\hat{p}_i$ is the confidence of sample i given by the softmax output of the model. The predicted accuracy for a given bin B_m is estimated from the proportion of correctly classified samples in the bin:

$$acc(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} 1(\hat{y}_i = y_i). \tag{15}$$

where $\hat{y}_i$ is the model's predicted classification and y is the true classification.

The confidence of a well-calibrated model is representative of it's true accuracy. In this work the calibration of the models is quantified by the Expected Calibration Error (ECE) and the more generalised Uncertainty Calibration Error (UCE). Based upon the work of Guo[19], the Expected Calibration Error quantifies the calibration of a model's softmax probabilities by taking a weighted average of the difference between the models' predicted accuracy and confidence for each bin:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)|. \tag{16}$$

Conversely the Uncertainty Calibration Error takes a weighted average of the difference between each bin's uncertainty and it's test error rate:

$$UCE = \sum_{m=1}^{M} \frac{|B_m|}{n} |err(B_m) - uncert(B_m)|. \tag{17}$$

Here $err(B_m)$ is the average fractional error for the data in a bin:

$$err(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} 1(\hat{y}_i \neq y_i), \tag{18}$$

where $uncert(B_m)$ is the average value of the uncertainty in bin m using one of the uncertainty metrics defined in sections 2.1-2.4. An error rate of 50% is expected for an untrained model. Any samples with $err > 50\%$ are assumed to be misclassified and $1 - err$ is used instead. Errors are then scaled by a factor of 2 such that an error rate of 1 corresponds to random decision-making. In this work, all UCE and ECE values are computed using 13 bins, each containing 8 samples and this is how the reliability diagrams are also plotted.

3.1. Spectral Normalisation

Spectral normalisation is a weight normalisation techinque primarily used in Generative Adversarial Networks (AGNs) that increases stability during network training[34]. A function, $f(x)$, is *Lipschitz continuous* if changing the input, x, cannot change it's output by more than the constraint $K \geq 0$ times this amount, i.e. the function's derivative is bounded[17]. The minimum value for K is called the *Lipschitz constant*, such that:

$$\text{Lip}_p(f) = sup \frac{\|f(x) - f(x')\|_p}{\|x - x'\|_p}. \tag{19}$$

If f is Lipschitz continuous and differentiable, the fraction can be rewritten as the Jacobian, $\|J_f(x)\|_p$[26]. If $f(x)$ is a linear map, such as a linear layer in a neural network, represented by a weight matrix, W,

then

$$\text{Lip}_p(f) = \|W\|_p = \sigma_{max}(W), \tag{20}$$

where $\sigma(W)$ can be obtained using the *singular value decomposition* (SVD) of W, which can be applied to any $m \times n$ matrix. In this work we used the built-in spectral normalisation Pytorch function with $n = 1$ power iterations to approximate the largest singular value[38]. For the models where spectral normalisation was applied, it was implemented in the final fully-connected layer.

4. E(2) Equivariant G-Steerable CNNs

G-steerable CNNs aim to preserve equivariance under a family of group symmetries, G. Transforming data, x, according to some transformation, $g \in G$, and passing it through a filter, κ, in the trained model must be equivalent to passing the data through the filter and transforming it:

$$\Phi(\kappa x) = \kappa' \Phi(x) \tag{21}$$

where $\Phi(x)$ is the output from the layer. For an equivariant filter, κ' is a linear representation of κ, i.e. $\kappa(gh) = \kappa(g)\kappa(h)$.

Following the method of Cohen & Welling[8], it can be shown that the standard convolution equation can be written in a generalised form for a group convolution:

$$[f * \kappa](g) = \sum_{h \in X} f(h)\kappa(g^{-1}h) \tag{22}$$

where $X = \mathbb{R}^2$ in the first layer and $X = G$ in further layers. This equation is in a more general form, however it is still only equivariant under translations. In order to introduce rotational equivariance, the filter kernel must be made rotation-steerable. Rotation-steerable kernels can be rotated and are constructed from a linear combination of basis functions. The type of transformation is identified by a representation of G, $\rho : G \to \mathbb{R}^{d \times d}$:

$$[g \cdot f](x) = \rho(g) f(g^{-1}x) \tag{23}$$

For E(2) equivariance to hold, the kernel must satisfy:

$$K(g \cdot x) = \rho_{out}(g) K(x) \rho_{in}(g^{-1}) \quad \forall g \in G, x \in \mathbb{R}^2, \tag{24}$$

where $K : \mathbb{R}^2 \to \mathbb{R}^{d_{in} \times d_{out}}$.

Planar images are defined by the input space $\mathbb{R}^2$. For single frequency and continuum radio images the feature field is scalar, such that $f : \mathbb{R}^2 \to \mathbb{R}^1$. The group representation for a scalar feature field is the trivial representation, $\rho(g) = 1$. The group representation of the output space can be chosen to suit a specific neural network model when building the network architecture.

In this work the G-steerable CNN of Weiler & Cesa has been implemented in which the Euclidian group is defined by the E(2), constructed to be the combination of the translation group and the orthogonal group $O(2)$: $E(2) \cong (\mathbb{R}, +) \rtimes O(2)$. The group therefore contains translations, continuous rotations and reflections. In this work we consider G-steerable CNNs which are equivariant under two subgroups of E(2), cyclic subgroups of the form $(\mathbb{R}, +) \rtimes C_N$, where C_N contains the set of discrete rotations $\{\frac{2\pi}{N}, \frac{4\pi}{N}, ..., 2\pi\}$, and the dihedral subgroups with the form $(\mathbb{R}, +) \rtimes D_N$, where $D_N \cong C_N \rtimes (\{\pm 1\}, *)$, which also includes reflections about the $x = 0$ axis. In this work only discrete rotations are considered of order N as convolution over a continous group is mathematically difficult to approximate, as is explained by Cohen & Welling[8].

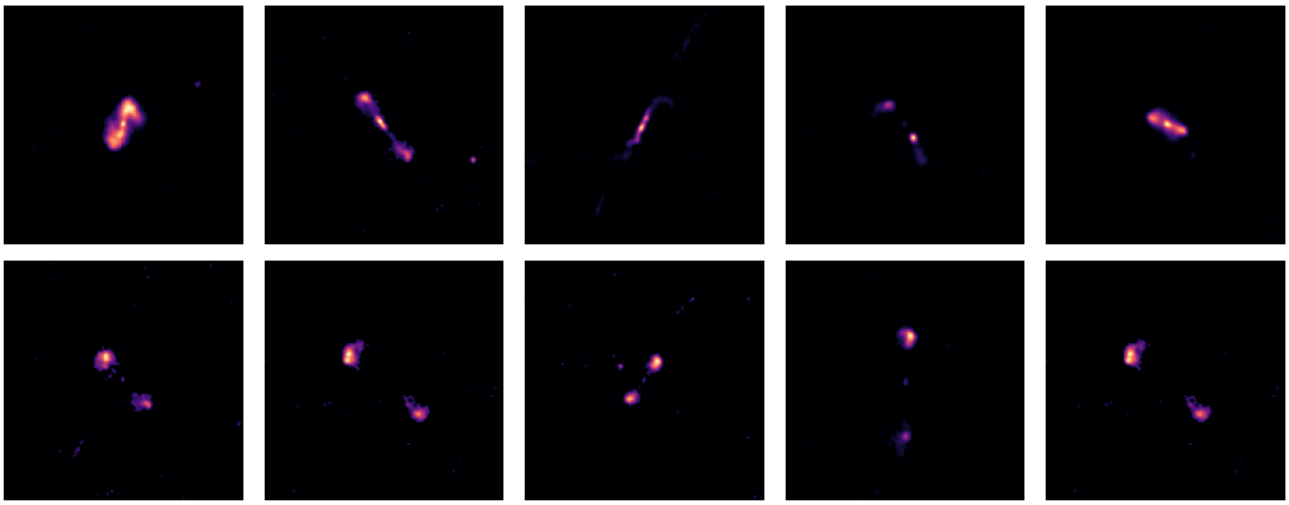

Figure 2: Input image data for 5 example FRI galaxies (top) and 5 example FRII galaxies (bottom) as classified in the *MiraBest** data set by the traditional *'by eye'* method.

5. Data

The data set used in this work is based upon the catalogue of Miraghaei & Best[33], who used a parent galaxy sample from Best & Heckman[4]. Best & Heckman cross-matched the Sloan Digital Sky Survey[47] data release 7[25] with the NRAO VLA Sky Survey[9] and the Faint Images of the Radio Sky at Twenty centimetres (FIRST)[3]. Miraghaei & Best morphologically classified the galaxies in the sample into FRI and FRII galaxies as defined by Fanaroff & Riley[12]. Sources were categorised using a 3-digit classification system as displayed in Table 1. Sources with an FRI morphology on one side and FRII on the other side were classified as *Hybrid*, while those which could not be visually classified were denoted as *Unclassifiable*. The second digit in the system is used to classify the human certainty in the decided classification, either being defined as *Confident* or *Uncertain*. The third digit represents any non-standard morphological phenomena including diffuse, 5 double-double, 53 wide-angle tail and 9 head-tail galaxies. Figure 2 displays example galaxies classified as FRI and FRII galaxies by the traditional *by eye* method.

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

Table 1: Miraghaei & Best[33] 3-digit classification system

Four pre-processing steps are applied to FITS images from the FIRST survey data in order to prepare the data for the machine learning process, according to the standard method followed by Aniyan & Thorat[1] and Tang et al.[42]:

(i) Image pixels with value below three times the local rms noise are set to zero value.

(ii) Images are cropped to 150×150 pixels.

(iii) All pixels outside a square region with extent equal to the furthest extent of the galaxy source from the image centre are set to zero value in order to remove any possible secondary background sources.

Operation	Kernel	Channels	Padding
Invariant Projection			
Convolution	5×5	6	1
ReLU			
Max-pool	2×2		
Convolution	5×5	16	1
ReLU			
Max-pool	2×2		
Invariant Projection			
Global Average Pool			
Fully-connected		120	
ReLU			
Fully-connected		84	
ReLU			
Dropout (p = 0.5)			
Fully-connected		2	

Table 2: The architecture for the LeNet-5 style network used in this work. The steps in italics are only applied for the G-Steerable models. The G-Steerable models also replace the standard convolutional layers with a group-equivariant convolutional layer.

(iv) The image is normalised according to:

$$\text{Output} = 255 \cdot \frac{\text{Input} - \min(\text{Input})}{\max(\text{Input}) - \min(\text{Input})}, \tag{25}$$

where 'Input' refers to the input image, 'Output' to the normalised output image and min(Input) and max(Input) refer to the minimum and maximum pixel values of the input image.

73 of the 1329 images in the Miraghaei & Best catalogue were removed to create the *MiraBest* data set, including: (i) 40 *unclassifiable* sources, (ii) 28 objects with extent greater than the 150×150 pixel boundary, (iii) 4 objects which partly overlapped the edge of the area covered by the FIRST survey and (iv) one object in the classification 103, which was removed as a minimum of two objects of a given classification are required to create the training:test data split. In this work the images classified as *Uncertain* are excluded aswell as those classified as *Hybrid* and the third digit in the classification system is ignored. This provides a binary classification subset of the *MiraBest* data set which is defined as the *MiraBest** data set.

6. Model

6.1. Architecture

This work implemented a standard LeNet-5 style neural network based upon the work of LeCun et al[31]. The network contains two convolutional layers and three fully connected layers. ReLU activation functions are applied at every layer and a max-pooling function is applied after each convolutional layer. A 50% dropout is applied at the final fully-connected layer as described by Krizhevsky[28] in order to approximate a Bayesian neural network. The full architecture of the model is displayed in Table 2.

The trivial representation is used for the input data. The choice of representation for all further layers is the regular representation, a common choice for finite groups such as C_N, D_N. The dimensionality of each group's representation space, $\mathbb{R}^{|G|}$, is equal to the order of the group, i.e. $\mathbb{R}^N$ for C_N and $\mathbb{R}^{2N}$ for

D_N. The regular representation acts by permuting the axes of the representation space. This repesentation is commonly used because it only permutes channels of fields and therefore is equivariant under ReLU activation functions aswell as max and average pooling functions as these are pointwise operations.

The e2cnn Pytorch extension[46] is used in this work for the G-Steerable models and the convolutional layers are replaced by a group-equivariant version. The group-equivariant convolutional layer is incompatible with the conventional fully-connected layers as the feature data is stored as a geometric tensor. Two additional steps must therefore be inserted into the G-steerable networks. It is necessary to reproject the tensor into a standard tensor format and then pooling is applied over the group features.

6.2. Training

The *MiraBest** data set is split by pre-defined training and test data partitions. The training partition is split further by a 80:20 split into training and validation sets. The data set is split into 584 training objects, 104 test objects and 145 validation objects.

Each network is trained over 600 epochs using a cross-entropy loss function and the Adam optimiser introduced by Kingma & Ba[27]. A batch size of 50 is used in each model. An initial learning rate of 10^{-4} and weight decay of 10^{-6} are used. A scheduler is implemented which reduces the value of the learning rate by 10% whenever the validation loss does not decrease over two consecutive epochs. An early-stopping criterion is implemented based on the model's validation accuracy and the model is saved when the early-stopping criterion is met.

Following training, the output predictions are compared against the target label and the percentage of incorrectly classified galaxies is calculated and used as the test error. Various metrics of uncertainty are used, as outlined in Section 2.

7. Results

7.1. Convergence and performance of G-Steerable CNNs

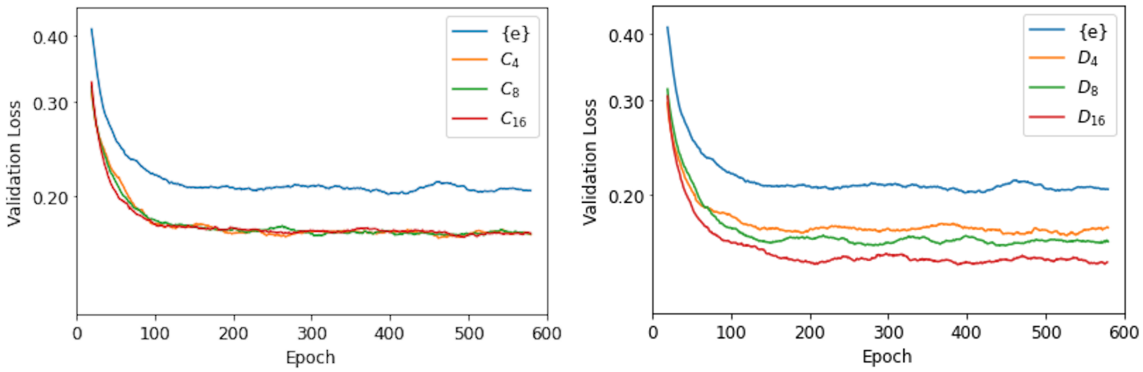

Figure 3: Validation losses measured over 600 epochs during the training of the standard LeNet-5 network, denoted {e} and (i) C_N group-equivariant CNN models (left) (ii) D_N group-equivariant CNN models (right) acting on the *MiraBest** data set.

We measure the validation loss over each epoch during training for each model and these values are plotted in Figure 3. The curves decrease to a point of stability at which point the model is saved. It can be seen that all forms of group-equivariant NN models show a considerable decrease in loss compared to the standard LeNet-5 model. The cyclic models show no difference in validation loss between orders, whereas the dihedral models show a considerable decrease in loss with increasing order. The higher order

models minimise the validation loss value faster in both the cyclic and dihedral case. This is in agreement with past works, such as that of Scaife & Porter[40] and Weiler & Cesa[46], who found the same for the MNIST dataset and attributed it to the improved ability of equivariant models to generalise better. It can be seen that, of the models considered, the dihedral $N = 16$ model had the minimum equilibrium validation loss.

	Original			Spectral Normalisation		
Model	FR I	FR II	Total	FR I	FR II	Total
$\{e\}$	93.47	94.27	93.90	93.83	95.00	94.45
C_4	94.78	94.06	94.40	**95.79**	94.09	94.89
C_8	**96.14**	93.75	94.88	95.34	95.78	95.57
C_{16}	95.91	94.56	95.19	95.29	95.66	95.49
D_4	95.91	**97.51**	**96.54**	95.51	96.94	96.27
D_8	95.16	94.10	94.60	95.66	**96.98**	**96.36**
D_{16}	95.51	95.79	95.65	95.52	94.29	94.87

Table 3: Model accuracy as a percentage for a standard LeNet-5 style model, denoted $\{e\}$ and it's group-equivariant equivalents. The values are taken for a model acting on a test set from the average value of $(1 - \text{error})$ over the test sample set. Accuracy is displayed separately for FRI and FRII classifications aswell as averaged over both classifications and for both with and without spectral normalisation applied at the final layer. The optimal model test accuracy is highlighted in bold for models with and withough spectral normalisation applied for the FRI and FRII sample subsets and also for the full sample set.

The accuracy is calculated for predictions of each sample on an untrained test data set. The test accuracy calculations are displayed in Table **??**. From the results it can be seen that the performance of the group-equivariant models is signficantly and consistently better than the standard LeNet model. The overall optimal test accuracy was achieved by the dihedral model with $N = 8$, which achieved an average test accuracy of 96.36%. It is noted that the dihedral models generally outperformed the cyclic models in FRII classification but both performed similarly in FRI classification. The model which achieved the highest test accuracy for FRI galaxy classification was the uncalibrated cyclic model with order $N = 8$, achieving an accuracy of 96.14%, while for FRII galaxy classification it was the uncalibrated dihedral model with order $N = 4$, achieving an accuracy of 97.51%.

	Confidence						Uncertainty $\langle\eta\rangle$					
Impact	**C4**	**C8**	**C16**	**D4**	**D8**	**D16**	**C4**	**C8**	**C16**	**D4**	**D8**	**D16**
Improved	37.50	22.12	25.00	43.27	34.62	33.65	20.19	17.31	15.38	21.15	19.23	17.31
Worsened	10.58	32.70	30.77	13.46	15.38	23.08	6.73	15.38	9.62	6.73	8.65	7.70

Table 4: The impact of applying group-equivariance for cyclic and dihedral subgroups of orders $N = \{4, 8, 16\}$ on the estimated values for confidence and uncertainty, where uncertainty is given by the average overlap as calculated using the method in Section 2.4. The impact for each model gives the percentage of objects which improved, i.e. $\langle\eta_{\{e\}}\rangle - \langle\eta_{G_N}\rangle > 0.01$, or worsened, i.e. $\langle\eta_{\{e\}}\rangle - \langle\eta_{G_N}\rangle < 0.01$, in model confidence following implementation of $G \in \{C, D\}, N \in \{4, 8, 16\}$ cyclic and dihedral G-steerable CNNs. All samples not accounted for showed no significant change in model confidence, i.e. $|\langle\eta_{\{e\}}\rangle - \langle\eta_{G_N}\rangle| < 0.01$.

We evaluate various uncertainty metrics for the numerous G-steerable CNN models used in this work and compare how the model uncertainty is impacted for two metrics, model confidence and overlap index, by implementing various degrees of group equivariance in Table 4. From the results it is shown that all models see a decrease in uncertainty for a larger proportion of the data set than see an increase in uncertainty. This improvement is also seen using the confidence metric for all models except for the cyclic models with orders $N = 8, 16$. It can be seen that the proportion of samples for which an improvement

is seen is considerably greater for dihedral models than for cyclic models and the proportion of samples where uncertainty worsens is lower for dihedral models except in the case of the $N = 4$ models. The general trend for model order is that the proportion of improved samples increases and the proportion of worsened samples decreases as order model decreases. Overall both the confidence and overlap index show the dihedral model with order $N = 4$ to have the greatest improvement in model confidence.

7.2. Uncertainty calibration for G-Steerable CNNs

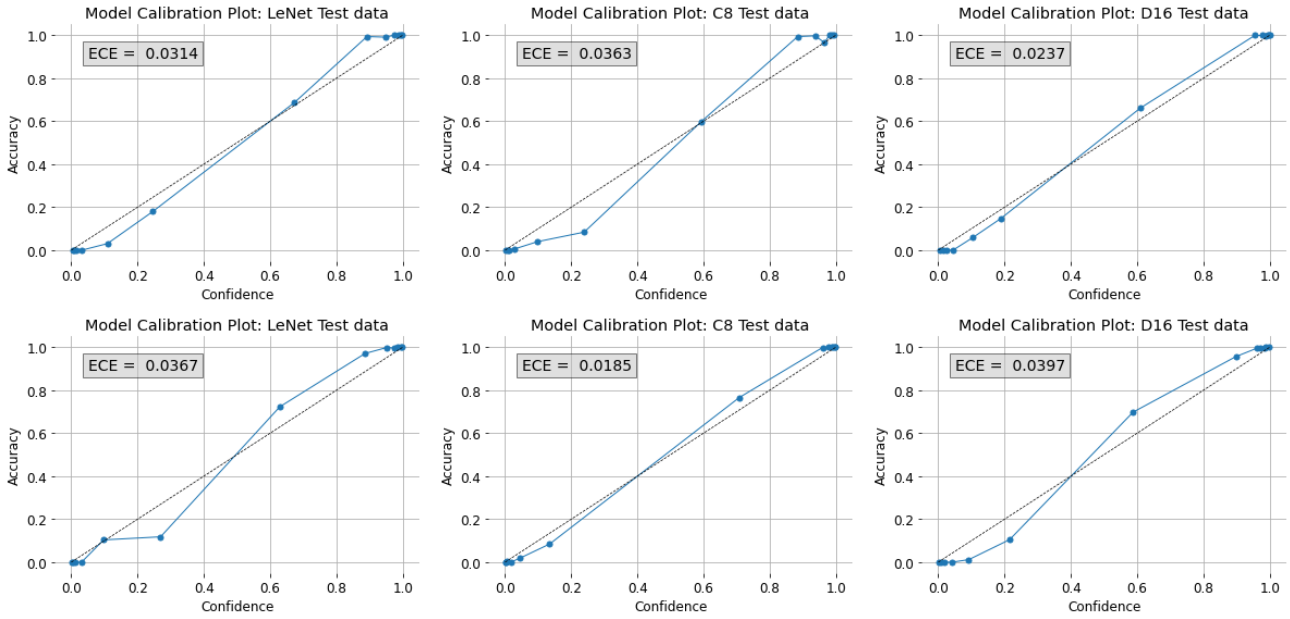

Figure 4: Reliability diagrams for a standard LeNet-5 style CNN (left), a C_8 group-equivariant CNN model (middle) and a D_{16} group-equivariant CNN model (right). Diagrams are displayed for both an unadapted model (top) and a model with spectral normalisation applied at the final layer (bottom). Confidence is given by the softmax output, while accuracy is given by the proportion of correctly classified objects. A perfectly calibrated network has no difference between confidence and accuracy at all confidence levels, represented by a dotted line. The ECE value is calculated according to the method in Section 2.5 and is displayed in the top left corner.

Figure 4 displays reliability diagrams of model test accuracy as a function of confidence for the LeNet neural network model and it's group-equivariant equivalents. Applied to the test data set, the accuracy was calculated as the proportion of FRII classifications using MC dropout over 100 stochastic runs. A test accuracy of 0 is therefore representative of a set comprising only FRI galaxies and a test accuracy of 1 of a test set comprising only FRII galaxies. The confidence was calculated as the softmax probability output for an FRII classification. A confidence level of 0.5 is therefore representative of 0% certainty, i.e. random decision-making. Confidence levels of 0 and 1 are representative of 100% certainty of FRI and FRII respectively. Therefore, if predictions of FRI and FRII galaxy classifications were equally well-calibrated, the graphs would be rotationally symmetric about $(0.5, 0.5)$. Considering the ECE values in Table **??** and the plots in Figure 4, the networks are shown to be relatively well-calibrated for all orders of group equivariance as they lie close to the identity line. However, the plots show a consistent trend of underconfidence. This can be seen as the plot primarily lies above the identity line for confidence values greater than 0.5 and under the line for confidence values less than 0.5. The lowest ECE value was achieved using the cyclic model with order $N = 8$ with spectral normalisation applied at the final layer. This model gave an ECE value of ECE $= 1.85\%$. Despite the lowest ECE value being achieved for a model with spectral normalisation applied, considering the results in Table **??** it can be seen that there

was no consistency to variations in ECE values following the implementation of spectral normalisation and therefore spectral normalisation had no consistent effect on how well-calibrated the model was using these metrics.

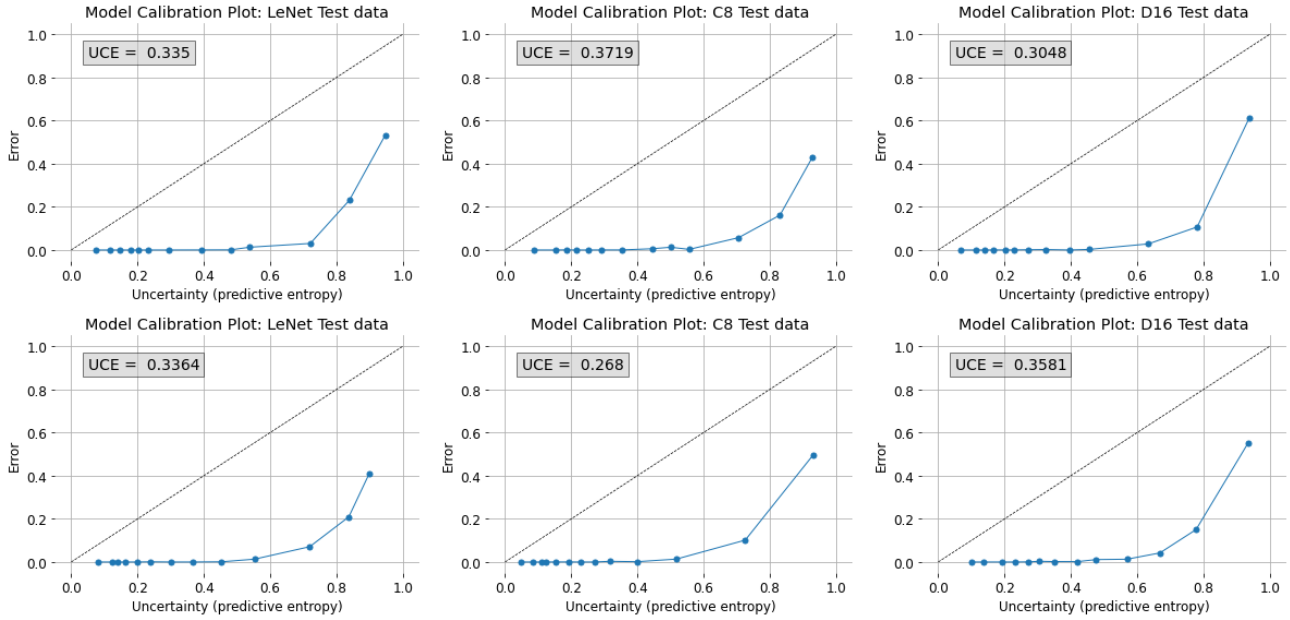

Figure 5: Reliability diagrams for a standard LeNet-5 style CNN (left), a C_8 group-equivariant CNN model (middle) and a D_{16} group-equivariant CNN model (right). Diagrams are displayed for both an unadapted model (top) and a model with spectral normalisation applied at the final layer (bottom). Uncertainty is given by the predictive entropy following the method outlined in Section 2.1, while error is given by the proportion of misclassified objects, as defined in Section 2.5. A perfectly calibrated network has no difference between uncertainty and error at all uncertainty levels, represented by a dotted line. The UCE value is calculated according to the method in Section 2.5 and is displayed in the top left corner.

Figure 5 displays reliability diagrams for the error rate of predictions on a test data set as a function of uncertainty with predictive entropy, $\mathbb{H}$, used as an uncertainty metric for the standard LeNet model, the C_8 model and the D_{16} model used in this work. The x axis is also scaled by $\ln 2$ to scale the entropy to a maximum value of 1. Using predictive entropy as an uncertainty metric, our original LeNet model and it's group-equivariant equivalents are all highly under-confident and very poorly calibrated. The models are much better calibrated as uncertainty approaches the boundaries 0 and 1. Using the predictive entropy as a metric of uncertainty, a minimum UCE value was achieved for the cyclic model with $N = 8$, with a calculated value of UCE $= 26.80\%$. Considering the values in Table **??** it can be seen that there is not a consistent difference seen in UCE values when spectral normalisation is applied, however the mean UCE value over all 7 orders improves slightly from 31.51% to 30.67%. Despite the lowest UCE value using predictive entropy as an uncertainty metric being achieved for a model with spectral normalisation applied, considering the results in Table **??** it can be seen that there was no consistent trend found in variations in the calibration of the models following the implementation of spectral normalisation.

Figure 6 displays reliability diagrams of the percentage error of predictions on a test data set as a function of uncertainty with overlap index, $\langle \eta \rangle$, used as an uncertainty metric. Using the overlap index as an uncertainty metric it can be seen that the models are all relatively well-calibrated. In the graphs displayed in Figure 6 it can be seen that the samples with low uncertainty lie below the identity line and for higher uncertainty values the points lie above the line. This is true for all the models considered and indicates that the model is under-confident at low uncertainty levels and becomes overconfident for higher uncertainty values. The point on the identity line at which the model is predicted to become

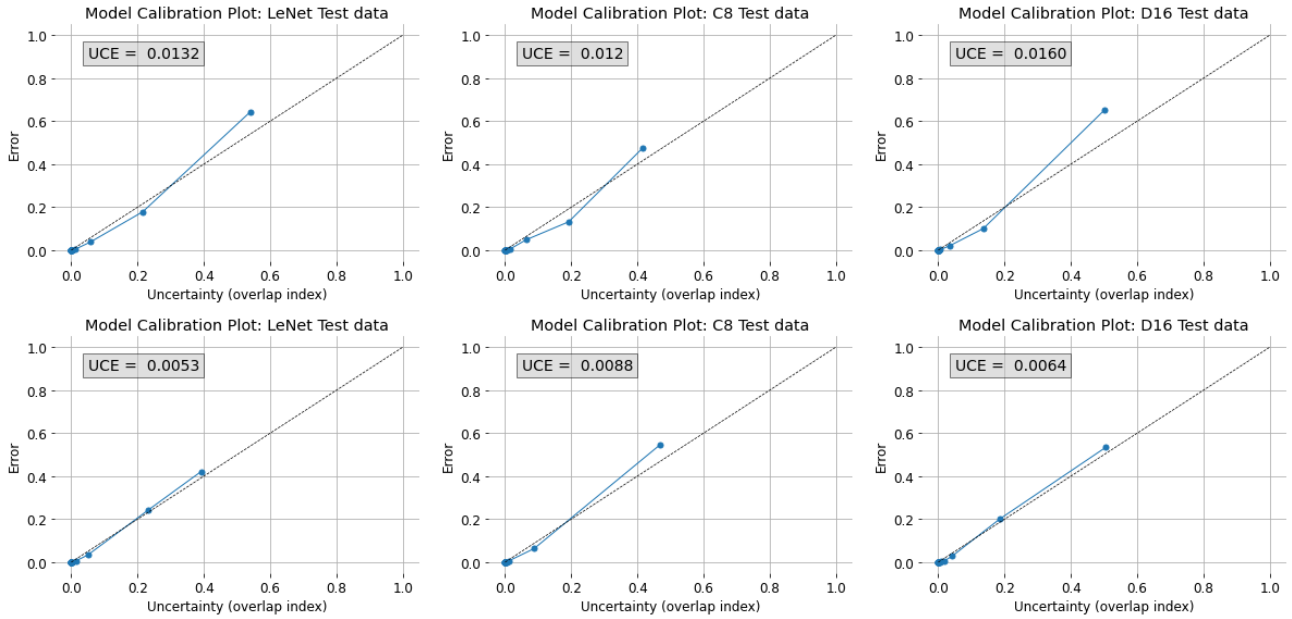

Figure 6: Reliability diagrams for a standard LeNet-5 style CNN (left), a C_8 group-equivariant CNN model (middle) and a D_{16} group-equivariant CNN model (right). Diagrams are displayed for both an unadapted model (top) and a model with spectral normalisation applied at the final layer (bottom). Uncertainty is given by the overlap index following the method outlined in Section 2.4, while error is given by the proportion of misclassified objects, as defined in Section 2.5. A perfectly calibrated network has no difference between uncertainty and error at all uncertainty levels, represented by a dotted line. The UCE value is calculated according to the method in Section 2.5 and is displayed in the top left corner.

overconfident is between 0.1 and 0.3 for all models considered. Using an overlap index as the uncertainty metric provided a minimum UCE values of 0.53%. This was achieved using two models, the cyclic model with order $N = 4$ with no spectral normalisation applied and the standard LeNet model with spectral normalisation applied. Despite one of the lowest UCE values using an overlap index as an uncertainty metric being achieved for a model with spectral normalisation applied, considering the results in Table **??** it can be seen that there was no consistent trend found in variations in the calibration of the models following the implementation of spectral normalisation.

	Original			Spectral Normalisation		
Model	**ECE** [%]	**UCE** $\langle\eta\rangle$ [%]	**UCE** $\mathbb{H}$ [%]	**ECE** [%]	**UCE** $\langle\eta\rangle$ [%]	**UCE** $\mathbb{H}$ [%]
$\{e\}$	5.46	1.32	33.50	3.67	**0.53**	33.64
C_4	2.81	**0.53**	30.95	3.37	1.34	33.33
C_8	3.63	1.20	37.19	**1.85**	0.88	**26.80**
C_{16}	2.43	0.68	29.40	3.55	0.84	28.73
D_4	2.98	0.67	**28.31**	2.83	1.03	28.07
D_8	2.91	0.80	30.76	2.58	1.07	28.37
D_{16}	**2.37**	1.60	30.45	3.97	0.64	35.81

Table 5: ECE and UCE values expressed as a percentage, calculated for a standard LeNet-5 style CNN, denoted as $\{e\}$, and cyclic and dihedral group-equivariant models of various orders. Values for the UCE are displayed for two different methods, one using an overlap index, $\langle\eta\rangle$, and one using the predictive entropy, $\mathbb{H}$. Values for all UCE and ECE values are taken for an unaltered model aswell as a model with spectral normalisation applied at the final layer. The minimum value for each ECE, UCE calculation is highlighted in bold.

$\text{ece}_u ce_t able$

Considering Table **??** it can be seen that the minimum UCE value achieved using the overlap index is ~ 50 times lower than the minimum UCE derived from predictive entropy and ~ 3 times greater than the minimum ECE value. The overlap index is therefore a considerably better-calibrated metric than predictive entropy or sample confidence for deducing model uncertainty and it produces uncertainties which are more representative of the true probability.

Predictions were made using the same models two additional times to determine the uncertainty on the ECE arising from the dropout. The standard deviation was found to be $\simeq 1\%$ for all models. This suggests that $N = 100$ stochastic runs is sufficient to produce a precise ECE value. The standard LeNet model was then retrained twice using refreshed initial weights and the standard deviation increased in value to $\simeq 6\%$.

8. Discussion

8.1. Analysis of uncertainty metrics

In Section 6.2 various uncertainty metrics were evaluated on the *MiraBest** data set and it was concluded that the overlap index is the best-calibrated uncertainty metric. Predictive entropy provides a metric for the distributive spread of softmax values for a given sample over repeated stochastic MC dropout runs. The outcome of this is that the softmax data observed can be spread out while the majority of the values predict the correct target classification. In this case the error rate would be much lower than the value for the predictive entropy and the metric would be a poor representation of model uncertainty. Conversely the overlap index measures the overlap of the posterior distribution and the alternative classifications through the softmax distributions. This is visualised for a well-defined classification with overlap $\langle \eta \rangle < 0.01$ and an uncertain classification with $\langle \eta \rangle = 0.51$ in Figure 1. This provides a better indication of how likely the model is to misclassify an object and therefore the uncertainty is better calibrated. The work of Niculescu-Mizil[36] found that neural networks using binary classifications are much better-calibrated than more complex classification systems. It was later discovered by Guo et al.[19] that modern, more complex networks with more layers are poorly calibrated compared to their simpler counterparts. It is therefore expected that the model used in this work would be relatively well-calibrated as the LeNet-5 network used is relatively simple and classifies galaxies by a binary classification system.

8.2. Analysis of model adaptations

By considering the results in Section 6.1, it can be seen that replacing the convolutional layers of a LeNet-5 style neural network with a group-equivariant equivalent produces a significant improvement in the accuracy of the model on an untrained test set. While it was shown that the uncalibrated dihedral model with order $N = 4$ produced results with the highest test accuracy, there is not a consistent systematic trend. It is also noted that it would seem logical for a truly rotationally invariant data set to converge to a maximum value for increased rotational freedom. It is therefore likely that these accuracy values may not represent an optimisation of the model as the variation is dominated by the high uncertainties. Similar findings were reached by Scaife & Porter[40], who concluded that the observations represented a limitation due to discretisation errors arising from rotating convolution kernels with small support. This is supported by the disparity between the results in this work and that of Scaife & Porter, in which the optimal model was found to be a dihedral model of order $N = 16$.

8.3. Notes on sample selection

One limiting factor in this work's analysis of model uncertainty metrics and their calibration is the lack of results for in-distribution uncertain data samples. As this work considers only the *Confident* MiraBest subset, we find that the results do not represent the full spectrum of real-world samples. The

repercussions of this can be seen in Figure 3 as most of the results have confidence values bunched near the boundaries at 0 and 1 and very few are seen for confidence values near 0.5. For the reliability diagrams in Figure 4 and 5 this is visualised by a lack of values with error greater than 0.5. Despite the lack of *Uncertain* samples in the sample data the predictive entropy outputs uncertainty values close to 1. This indicates that the metric is a poor representation of the model confidence. While the overlap index proves to be relatively well-calibrated overall for the samples in the subset, from Figure 6 it can be seen that the model becomes less well-calibrated for more uncertain samples. It is possible that this trend would continue for more uncertain samples and therefore the model would be less well-calibrated for real-world samples than the results suggest as samples are selectively curated. Considering *Uncertain* in-distribution data in the test data set would provide insight into the ability of the network at higher uncertainties and it's efficacy when applied to a real-world data set, aswell as providing a basis for possible recalibration. In future models it could be important to introduce a third classification for out-of-distribution *Hybrid* data samples in order to further generalise the application of the model to correctly classify new data sets.

8.4. Notes on standout misclassification

In Figure 7 an example galaxy is shown which is incorrectly classified by the network with a high confidence. The galaxy has been classified using traditional methods as an FRII galaxy, however the model classifies it as FRI. The uncertainty value is found to be low for all uncertainty metrics and therefore the model's confidence in it's prediction is generally very high. The galaxy is misclassified due to particularly bright components close to the extremities of the source.

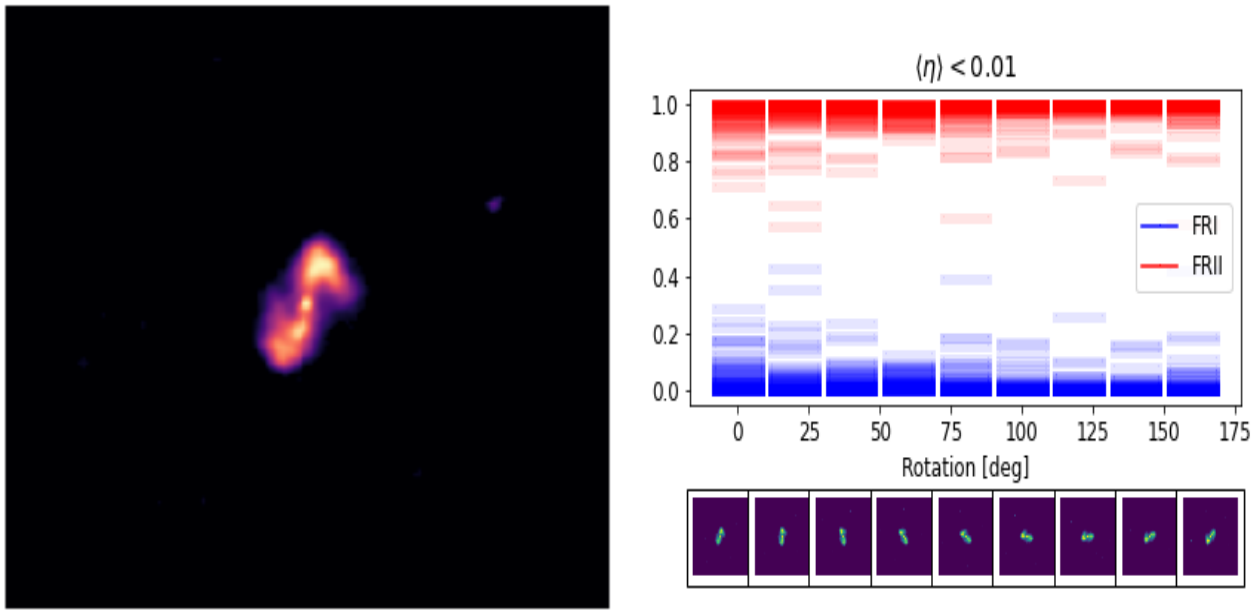

Figure 7: The *MiraBest** image of a galaxy that has been misclassified with high confidence (left) and the corresponding softmax outputs for 100 forward passes through the D_8 group-equivariant CNN (right). The mean overlap is displayed above the diagram of softmax outputs and shows a low softmax value and therefore high prediction confidence, however the galaxy is misclassified.

8.5. Criticism of MC Dropout as Bayesian approximation

This work relies on the assumption that the MC dropout method described in Section 2 is a valid method for producing a predictive posterior distribution. However, work by Le Folgoc et al.[30] suggests that the implementation of the MC dropout method is actually a poor approximation of Bayesian

computation as it fails to represent the true posterior distribution. This appears to support our calculated standard deviation value for the ECE derived in Section 7.2. This result appears to suggest that the MC dropout method applied in this work is useful to provide rough estimates on the relative calibration of different uncertainty metrics, however it does not provide precise results and therefore does not provide a useful comparison between the varying orders of group equivariance for the various G-steerable models analysed. An additional issue with the ECE as a calibration metric is it's high sensitivity to the number of bins used in the calculation[29]. The UCE does not face this issue and therefore may be a better representation of the true uncertainty calibration for a model.

9. Conclusions

In this work we have presented the application of uncertainty calibration measurements to deep learning Fanaroff-Riley galaxy classification. Posterior uncertainties were calculated using numerous metrics based upon the use of the Monte Carlo dropout method as an approximate Bayesian computation to produce posterior probability distributions for predictions of the model.

We find that using an overlap index as an uncertainty metric produces a better predictive uncertainty calibration than the use of predictive entropy or model confidence, with a minimum UCE value ~ 50 times lower than using predictive entropy and ~ 3 times lower than the lowest ECE achieved using model confidence. This suggests that the overlap index is considerably more accurate as a metric for model uncertainty in this case.

Our analysis of different uncertainty metrics indicates that the reason that the overlap produces better uncertainty calibration results is that the predictive entropy is representative of the width of a prediction's posterior distribution which is not necessarily representative of the uncertainty on a given class prediction, whereas the overlap index is a measure of the separation between the posterior distribution and alternative class predictions.

We have considered the calibration of a standard LeNet-5 style CNN and various equivalent CNNs with different orders of group-equivariance. We find that all group-equivariant models achieve a greater average accuracy than the standard LeNet model. We also presented the first application of spectral normalisation to deep learning Fanaroff-Riley classification. Spectral normalisation was applied at the final fully-connected layer of the network to observe the impact on the model uncertainty calibration.

Using the overlap index as an uncertainty metric the minimum UCE value is achieved for the unmodified C_4 model and the standard LeNet model with spectral normalisation applied. This indicates that the model is better calibrated for simpler models, such as those with a lower order of rotational equivariance and those which are limited to cyclic equivariance without reflectional equivariance. This is consistent with expectations that simple neural networks tend to be better calibrated, however overall there is not a consistent improvement in uncertainty calibration across all the models. It is also found that there is a significant variation in uncertainty calibration metric values when the LeNet model is retrained, with an observed variation in the ECE value of $\sim 6\%$. A criticism of the Monte Carlo dropout method is considered as a reasoning for the uncertainty seen when retraining the models. It is recommended that future work considers alternative methods to produce a more precise Bayesian posterior distribution.

Our analysis of the implementation of spectral normalisation is that there were no significant or consistent changes to the uncertainty calibration of the network. It is noted that the high level of uncertainty that is derived from the training of the model may obscure any underlying consistent patterns which would be observed.

In this work we have used a binary morphological classification system, a simplification of the real population of radio galaxies. We have also curated the *MiraBest* data set to exclude uncertain data. Our analysis of our results using the overlap index as an uncertainty metric concluded that results based on

our test data do not represent the full spectrum of real-world data samples, leading to a gap in results at high uncertainty values. The effective application of these neural networks to real-world scenarios will be key in capitalising on the current growth in radio astronomy data collection rates and it is therefore essential that we work on generalising these models to include all classifications.

References

[1] ANIYAN, A. K., AND THORAT, K. Classifying radio galaxies with the convolutional neural network. The Astrophysical Journal Supplement Series 230, 2 (jun 2017), 20.

[2] BANFIELD, J. K. ET AL. Radio galaxy zoo: host galaxies and radio morphologies derived from visual inspection. Monthly Notices of the Royal Astronomical Society 453, 3 (09 2015), 2326–2340.

[3] BECKER, R. ET AL. The FIRST survey: Faint images of the radio sky at twenty centimeters. 450 (Sept. 1995), 559.

[4] BEST, P. N. AND HECKMAN, T. M. On the fundamental dichotomy in the local radio-AGN population: accretion, evolution and host galaxy properties. Monthly Notices of the Royal Astronomical Society 421, 2 (03 2012), 1569–1582.

[5] BOWLES, M., SCAIFE, A. M. M., PORTER, F., TANG, H., AND BASTIEN, D. J. Attention-gating for improved radio galaxy classification. 501, 3 (Mar. 2021), 4579–4595.

[6] CHEN, T., FOX, E., AND GUESTRIN, C. Stochastic gradient Hamiltonian Monte Carlo, 22–24 Jun 2014.

[7] CHEN, W. ET AL. Radio galaxy zoo: CLARAN – a deep learning classifier for radio morphologies. Monthly Notices of the Royal Astronomical Society 482, 1 (10 2018), 1211–1230.

[8] COHEN, T. S., AND WELLING, M. Group equivariant convolutional networks. arXiv e-prints (Feb. 2016), arXiv:1602.07576.

[9] CONDON, J. J. ET AL. The NRAO VLA sky survey. The Astronomical Journal 115, 5 (may 1998), 1693.

[10] DEGROOT, M. H., AND FIENBERG, S. E. The comparison and evaluation of forecasters. Journal of the Royal Statistical Society. Series D (The Statistician) 32, 1/2 (1983), 12–22.

[11] DIELEMAN, S., FAUW, J. D., AND KAVUKCUOGLU, K. Exploiting cyclic symmetry in convolutional neural networks. CoRR abs/1602.02660 (2016).

[12] FANAROFF, B. L., AND RILEY, J. M. The morphology of extragalactic radio sources of high and low luminosity. Monthly Notices of the Royal Astronomical Society 167, 1 (04 1974), 31P–36P.

[13] GAL, Y. Uncertainty in deep learning. PhD thesis, University of Cambridge.

[14] GAL, Y., AND GHAHRAMANI, Z. Bayesian convolutional neural networks with Bernoulli approximate variational inference, 2015.

[15] GAL, Y. AND GHAHRAMANI, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning, 2015.

[16] GARRINGTON, S. T. ET AL. e-MERLIN. In Ground-based Telescopes (2004), J. M. O. Jr., Ed., vol. 5489, International Society for Optics and Photonics, SPIE, pp. 332 – 343.

[17] GOUK, H., FRANK, E., PFAHRINGER, B., AND CREE, M. J. Regularisation of neural networks by enforcing Lipschitz continuity. Mach. Learn. 110, 2 (feb 2021), 393–416.

[18] GRAVES, A. Practical variational inference for neural networks. In Advances in Neural Information Processing Systems (2011), vol. 24, Curran Associates, Inc.

[19] GUO, C. ET AL. On calibration of modern neural networks. CoRR abs/1706.04599 (2017).

[20] HÜLLERMEIER, E., AND WAEGEMAN, W. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Machine Learning 110, 3 (2021), 457–506.

[21] JARVIS, M. ET AL. The MeerKAT international GHz tiered extragalactic exploration (MIGHTEE) survey. PoS MeerKAT2016 (2018), 006.

[22] JIA, X. ET AL. SPINBIS: Spintronics-based Bayesian inference system with stochastic computing. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 39, 4 (2020), 789–802.

[23] JIA, X. ET AL. Efficient computation reduction in Bayesian neural networks through feature decomposition and memorization. IEEE Transactions on Neural Networks and Learning Systems 32, 4 (2021), 1703–1712.

[24] JOHNSTON, S. ET AL. Science with ASKAP. the Australian square-kilometre-array pathfinder. Experimental Astronomy 22, 3 (Dec. 2008), 151–273.

[25] KEVORK, N. ET AL. The seventh data release of the Sloan digital sky survey. The Astrophysical Journal Supplement Series 182, 2 (may 2009), 543.

[26] KIM, H., PAPAMAKARIOS, G., AND MNIH, A. The Lipschitz constant of self-attention. In Proceedings of the 38th International Conference on Machine Learning (18–24 Jul 2021), M. Meila and T. Zhang, Eds., vol. 139 of Proceedings of Machine Learning Research, PMLR, pp. 5562–5571.

[27] KINGMA, D., AND BA, J. Adam: A method for stochastic optimization. International Conference on Learning Representations (12 2014).

[28] KRIZHEVSKY, A. ET AL. ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems (2012), vol. 25, Curran Associates, Inc.

[29] LAVES, M., IHLER, S., KORTMANN, K., AND ORTMAIER, T. Well-calibrated model uncertainty with temperature scaling for dropout variational inference. CoRR abs/1909.13550 (2019).

[30] LE FOLGOC, L. ET AL. Is MC dropout Bayesian? CoRR abs/2110.04286 (2021).

[31] LECUN, Y., BOTTOU, L., BENGIO, Y. AND HAFFNER, P. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 11 (1998), 2278–2324.

[32] LUKIC, V. ET AL. Radio galaxy zoo: compact and extended radio source classification with deep learning. Monthly Notices of the Royal Astronomical Society 476, 1 (01 2018), 246–260.

[33] MIRAGHAEI, H. AND BEST, P. N. The nuclear properties and extended morphologies of powerful radio galaxies: the roles of host galaxy and environment. Monthly Notices of the Royal Astronomical Society 466, 4 (01 2017), 4346–4363.

[34] MIYATO, T ET AL. Spectral normalization for generative adversarial networks, 2018.

[35] MOHAN, D., SCAIFE, A. M. M., PORTER, F., WALMSLEY, M., AND BOWLES, M. Quantifying uncertainty in deep learning approaches to radio galaxy classification. Monthly Notices of the Royal Astronomical Society 511, 3 (jan 2022), 3722–3740.

[36] NICULESCU-MIZIL, A., AND CARUANA, R. Predicting good probabilities with supervised learning. ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning (01 2005), 625–632.

[37] O'SHEA, K., AND NASH, R. An introduction to convolutional neural networks. CoRR abs/1511.08458 (2015).

[38] PAN, Z., AND MISHRA, P. Fast approximate spectral normalization for robust deep neural networks. CoRR abs/2103.13815 (2021).

[39] PASTORE, M. AND CALCAGNÌ, A. Measuring distribution similarities between samples: A distribution-free overlapping index. Frontiers in Psychology 10 (2019).

[40] SCAIFE, A. M. M., AND PORTER, F. Fanaroff–Riley classification of radio galaxies using group-equivariant convolutional neural networks. Monthly Notices of the Royal Astronomical Society 503, 2 (02 2021), 2369–2379.

[41] SHIJIE, J. ET AL. Research on data augmentation for image classification based on convolution neural networks. In 2017 Chinese Automation Congress (CAC) (2017), pp. 4165–4170.

[42] TANG, H. ET AL. Transfer learning for radio galaxy classification. Monthly Notices of the Royal Astronomical Society 488, 3 (07 2019), 3358–3375.

[43] TICKNOR, J. L. A Bayesian regularized artificial neural network for stock market forecasting. Expert Systems With Applications 40, 14 (2013), 5501–5506.

[44] VAN HAARLEM, M. P. ET AL. LOFAR: The low-frequency array. A&A 556 (2013), A2.

[45] WANG, R., WALTERS, R., AND YU, R. Data augmentation vs. equivariant networks: A theory of generalization on dynamics forecasting, 2022.

[46] WEILER, M., AND CESA, G. General $E(2)$-equivariant steerable cnns. arXiv e-prints (Nov. 2019), arXiv:1911.08251.

[47] YORK, D.G. ET AL. The Sloan digital sky survey: Technical summary. The Astronomical Journal 120, 3 (sep 2000), 1579.