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

Samuel Hudson 10458093

Supervised by: Anna M. M. Scaife
Project partner: Alex Millicheap

Abstract

In order to correctly classify new galaxy images in the radio spectrum, accurate but also well-calibrated models are required. Monte Carlo dropout is used to approximate a probabilistic Bayesian model in order to extract probabilities and uncertainties on a binary Fanaroff-Riley classification task. Equivariant convolutional neural networks can provide highly accurate classifications of radio galaxies and in this work we show how calibrated these networks are using 2 calibration metrics, the ECE and UCE. We find all models to output reasonably well calibrated probabilities with an error on the order of 3%, and a slight under-confidence was also observed. It was found that using the overlap index as an uncertainty measure as opposed to the commonly used predictive entropy, provides calibration errors on the order of 1% with a best calibration error of 0.53%. We apply regularisation in the form of spectral normalisation and show that this has no significant effect on the calibration scores while noting that any trends may be obscured by large uncertainties. A potential case of label noise in the MiraBest Confident dataset is identified by applying the uncertainty metrics from our models.

1. Introduction

The recent development of new radio telescopes such as the Australian Square Kilometre Array Pathfinder (ASKAP) [McConnell et al. 2020] and MeerKAT [Jarvis et al. 2016] generate massive volumes of astronomical data that is increasingly difficult to group and classify by conventional means. In order to automate the classification of this data there has been a movement towards machine learning tools that can quickly identify and sort novel data after being trained on curated data-sets.

A widely used classification scheme in radio astronomy is the Fanaroff-Riley (FR) type [Fanaroff and Riley 1974] first described in 1974 for objects with an active galactic nucleus, which differentiates between two possible morphologies for a radio galaxy. FR I galaxies are identified by bright central jets that emanate from the supermassive black holes at their centre. FR II galaxies are characterised by bright termination shocks at the end of each of the jets. Both types of galaxy exist on a spectrum, meaning some ambiguity is present when deciding which category to place a galaxy into. This makes it crucial for any machine learning model to encapsulate this uncertainty rather than provide an under or over-confident answer that could result in incorrect classification. The flagging of uncertain sources allows for further inspection by humans on a much smaller set of galaxies while confidently classified galaxies will need no such intervention, hugely increasing the rate at which new labels can be assigned.

Unlike in other fields which use deep learning, radio astronomy has a significant lack of accurately labelled data. In the case of FR classification, the largest dataset available is the MiraBest dataset [Miraghaei and Best 2017] which contains only 1329 labelled images. Data sets used for image classification such as MNIST [Lecun et al. 1998] contain orders of magnitude more data. This makes any errors and biases in the data more prevalent in the trained models and makes over-fitting to the given data an issue. When using any kind of applied machine learning, it is important to identify and minimise these biases as this can affect the generalisation ability of a model when tackling new and unseen information. Over-fitting and under-fitting in models can occur when the training data does not sufficiently represent the true distribution of data, leading to poorly calibrated models that may report meaningless probabilities and uncertainties.

Classical convolutional neural networks (CNNs) are preferred over standard multi-layer perceptrons (MLPs) for image classical due to the learning of convolutional kernel weights during training, leading to feature extraction that is naturally translation invariant [Weiler and Cesa 2019]. However, conventional CNNs are not rotation and reflection equivariant, and are therefore prone to learning copies of a convolutional kernel in multiple orientations if the data has no preferred orientation/chirality. Augmentation in the form of random rotations and reflections of the input data can be applied during the training process to approximate equivariance but this does not guarantee that it will generalise to new data [Simard, Steinkraus, and J. Platt 2003]. These augmentations must be warranted, as unprincipled augmentation may lead to issues such as the cold posterior effect [Noci et al. 2021]. The advent of Group-Steerable CNNs [Cohen and Welling 2016] allows the necessary isometries to be encoded directly into the model and reduces the learning of redundant parameters which can lead to improvements in performance [Dieleman, De Fauw, and Kavukcuoglu 2016]. One important group is the Euclidean group E(2) which is the set of transformations that act on the 2D image plane $\mathbb{R}^2$ and cover all rotations, reflections, translations and combinations thereof. E(2)-Steerable CNNs [Weiler and Cesa 2019] have been shown to perform better than non-equivariant models by Scaife and Porter 2021 when applied to the problem of radio galaxy classification. We use and then modify similar models in this report to examine how well calibrated these model probabilities are, and decide on the best metric to encapsulate the uncertainty in a source.

Rather than providing a binary classification, Bayesian neural networks (BNNs) and their approximations [Gal and Ghahramani 2015] are designed to provide probabilistic outputs, hence the given probability should equate to the chance of misclassifying a given sample. It is also possible to extract uncertainties on the probabilities by considering the models output distribution, the most popular uncertainty metric is the entropy [Namdari and Li 2019], however alternatives are available such as the the distribution free overlap index [Pastore and Calcagnì 2019]. Both the probabilities and uncertainties should be representative of the chance of misclassifying a sample; calibration metrics are commonly used to quantify how well the model outputs correspond to their expected results [Nixon et al. 2019]. In practice a perfectly calibrated model is unattainable and calibration techniques may need to be applied for models to output meaningful probabilities and uncertainties. Calibration methods can be implemented by making changes directly to the model, requiring retraining, while other methods can be applied as a much more computationally cheap post-processing step such as Platt scaling [J. C. Platt 1999].

Contributions: This report will look at the calibration of the modified LeNet and E(2)-equivariant models used by Scaife and Porter 2021, specifically the models of order 4, 8 and 16. We show how the calibration of probabilities and uncertainties are affected by implementing equivariance and also find the most appropriate metric for quantifying the uncertainty in these models. Finally, we introduce spectral normalisation, a technique that is commonly applied to generative adversarial networks (GANs) as a training stabiliser [Miyato et al. 2018], as a potential calibration method and examine the effects, if any, on the calibration of both non-equivariant and cyclic/dihedral-equivariant models.

2. Uncertainty Quantification

2.1. Bayesian Neural Networks

In standard neural networks, the likelihood $\mathfrak{L}(D|w)$ of the dataset D given model parameters w is maximised during training in order to provide point-wise predictions of unseen data. A Bayesian Neural Network (BNN) aims to provide probabilistic outputs from a posterior distribution $P(w|D)$, obtained using Bayes theorem

$$P(w|D) = \frac{\mathfrak{L}(D|w)P(w)}{Z} \tag{1}$$

where $P(w)$ is a prior distribution of model weights before training, commonly assumed to be Gaussian , and $Z = P(D)$ is known as the evidence, which we assume to be a constant.

The predictive posterior distribution of a BNN is $P(D^*|D)$ for an unseen data point D^* and is obtained by integrating out the variational parameters, w, of the model

$$P(D^*|D) = \int P(D^*|w)P(w|D)dw \tag{2}$$

where $P(D^*|w)$ is the predictive distribution of the class of D^* given some model parameters w. The predictive posterior distribution is commonly determined through variational inference [Blei, Kucukelbir, and McAuliffe 2017], however we can approximate a BNN using Monte-Carlo (MC) dropout to integrate over the model weights by taking samples from the model as demonstrated by Gal and Ghahramani 2015. The predictive posterior is then given by

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

where $P(D^*|w^i)$ is the prediction of D^* given a subset of weights $w^i \subseteq w$. MC-dropout is used to determine w^i by turning off weights in the model at random using a Bernoulli process with probability $p_d = 0.5$.

A distribution of probabilities $\mathbf{q}(y = c|x, D)$ of class c is generated for an input image x giving model output y trained on dataset D by performing N forward passes of each test sample with MC-dropout enabled. Each sample $q_i(y_i = c|x, w^i)$ is obtained from the softmax function

$$q_i(y_i = c|x, w^i) = \frac{\mathrm{e}^{-\beta z_i}}{\sum_{i=1}^{K} \mathrm{e}^{-\beta z_i}} \tag{4}$$

where z_i are the components (commonly referred to as logits) of the raw output vector of size $K = 2$ in the case of binary classification and β is constant factor that determines the weighting of small outputs and can thus be tuned as a hyper-parameter to calibrate the probabilities, $1/\beta$ is commonly referred to as the 'Temperature' [Laves et al. 2019]. This normalises the output vector such that the sum of its components $\sum_{i=1}^{K} q_i = 1$. We can then use $\mathbf{q}$ to derive representations of the uncertainty, and the average softmax probability $\hat{q}$ for a given class is used as the confidence in the model prediction.

2.2. Overlap Index

One method to quantify the uncertainty in this output distribution is to use the overlap index $\langle\eta\rangle$ [Pastore and Calcagnì 2019], which is determined by first performing a calculation of the local densities at each position $z_i \in [0, 1]$ in steps of $\delta z = \frac{1}{M_z}$. This is done using a Gaussian kernel for each class in the output:

$$f_y(z) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\beta\sqrt{2\pi}} \mathrm{e}^{-(z-y_i)^2/2\beta^2} \tag{5}$$

where $\beta = 0.1$ is a predetermined constant, y is a class and y_i are the softmax outputs belonging to the given class. The overlap index is then calculated as

$$\langle\eta\rangle = \sum_{i=1}^{N} \min[f_{FRI}(z_i), f_{FRII}(z_i)]\delta z \tag{6}$$

and takes values in the range $[0, 1]$, with low values indicating high confidence and hence a distinct separation of classes in the model output. High values close to 1 suggest the model is uncertain about its output and there is a high degree of mixing in the results.

2.3. Entropy and Mutual Information

Alternatively, we can use the entropy of the softmax output distribution to quantify the uncertainty. The normalised predictive entropy, $\mathbb{H}$, of a distribution of N softmax probabilities obtained using MC-dropout can be approximated [Gal 2016] as

$$\mathbb{H}(y|x^*, D) = -\frac{1}{\ln(C)} \sum_{c}^{C} \left(\frac{1}{N} \sum_{i=1}^{N} q_i(y_i = c|x^*, w^i) \right) \ln \left(\frac{1}{N} \sum_{i=1}^{N} q_i(y_i = c|x^*, w^i) \right). \tag{7}$$

where we sum over all classes $c \in$ [FR I, FR II] for an unseen data point x^*. In order to use the predictive entropy as an uncertainty, it is scaled onto the range [0,1] by the natural logarithm of the total number of classes as shown in Equation 7. Predictive entropy encapsulates both aleatoric and epistemic uncertainties and is the sum of the mutual information, $\mathbb{I}$, a measure of the epistemic uncertainty, and average entropy, $\mathbb{E}_q\mathbb{H}$, a measure of aleatoric uncertainty for in-distribution data [Mukhoti et al. 2021]. The (normalised) average entropy can also be approximated using MC samples as

$$\mathbb{E}_q\mathbb{H}(y|x^*, D) = -\frac{1}{\ln(C)}\sum_{c}^{C}\frac{1}{N}\sum_{i=1}^{N}\big(q_i(y_i = c|x^*, w^i)\ln(q_i(y_i = c|x^*, w^i))\big) \tag{8}$$

which is the average of the entropy of each stochastic forward pass as opposed to the entropy of the average model prediction and is high for noisy/ambiguous samples. The average entropy is also normalised in a similar fashion to predictive entropy. The mutual information can then be expressed as the difference between the predictive and average entropy,

$$\mathbb{I}(y|x^*, D) = \mathbb{H}(y|x^*, D) - \mathbb{E}_q\mathbb{H}(y|x^*, D) \tag{9}$$

which quantifies the uncertainty inherent in the model and can be reduced with more training data.

2.4. Calibration Metrics

In a binary classifier, each test image returns an average softmax probability $\hat{q}$ and a predicted class label $\hat{y}$. In order to consider a model calibrated, it requires that the confidence in the classification is equal to the probability of classifying the image correctly

$$P(\hat{y} = c) = \hat{q} \tag{10}$$

where c is the true galaxy classification and $\hat{q}$ is the softmax probability which we use as our measure of confidence. The model outputs from the test data-set are placed into M bins with an equal number of samples in each bin. The average confidence and accuracy of each bin is then compared to quantify the degree of mis-calibration in the given bin. We define confidence in the model output for a bin B_m as the average softmax probability of a positive (FR II) classification

$$conf(B_m) = \frac{1}{|B_m|}\sum_{j\in B_m}\hat{q}_j(y_j = 1). \tag{11}$$

Accuracy is then defined as the proportion of positive classifications in a bin B_m

$$acc(B_m) = \frac{1}{|B_m|}\sum_{j\in B_m}\mathbf{1}(\hat{y}_j = 1) \tag{12}$$

where the function $\mathbf{1}(\hat{y}_j = 1)$ denotes the proportion of positive classifications from all stochastic forward passes of a given test image. This is done so that for a given test image, an accuracy of 1 corresponds to 100% classification as FR II, while an accuracy of 0 means 100% classification as FR I.

Expected Calibration Error (ECE) is then defined as

$$ECE = \sum_{m=1}^{M}\frac{|B_m|}{n}\,|acc(B_m) - conf(B_m)| \tag{13}$$

where n is the size of the dataset and this represents the % mis-calibration between the probability and true classification accuracy.

We quantify the accuracy in the model as the proportion of positive (FRII) classifications when performing $N = 100$ stochastic forward passes through the model using MC-dropout and $N_{rot} = 9$ discrete rotations of the test image by 20 degrees, then comparing the softmax output to the target classification for each pass.

We define the error to be the proportion of misclassifications from N forward passes through the model. We expect an untrained model to give a 50% error rate, and so all our error values are doubled such that an error of 1 corresponds to 'guessing', which should also correspond to the highest uncertainty. Any error rates $> 50\%$ are considered mis-classified images and may be due to over-fitting or in some cases label noise. Therefore we take the error to be $\min(err, 1 - err)$ such that the maximum error rate is 50%. We then double the error so that an error of 1 should correspond to maximum uncertainty (both overlap and entropy lie on the range [0,1]). Any confidently mis-classified samples were therefore treated as correctly classified for the purposes of calibrating the model, as we only desire the uncertainty metric to represent the degree of mixing between class outputs rather than penalise the model for possible label noise and/or poor accuracy which must be improved independently from the calibration.

The Uncertainty Calibration Error (UCE) is defined as

$$UCE = \sum_{m=1}^{M} \frac{|B_m|}{n} \left| uncert(B_m) - err(B_m) \right| \tag{14}$$

where $uncert(B_m)$ is the averaged uncertainty measure in the bin B_m (either overlap index or predictive entropy) and $err(B_m)$ is the average frequentist error in the bin B_m. Since classification error is calculated for every image in the test data, it is not necessary to use binning, instead each point can contribute to the UCE equally with $|B_m| = 1$.

2.5. E(2) Equivariant G-Steerable CNNs

Conventional CNNs are by design translation invariant, however when dealing with data that is expected to be generally independent of orientation and chirality/reflections, such as radio galaxies, the use of conventional CNNs can encourage models to learn unnecessary rotation and reflection dependant parameters. The first work done on equivariant CNNs was by Cohen and Welling 2016 and was restricted to 90 degree rotations and reflections. Weiler, Hamprecht, and Storath 2018 developed the necessary framework for steerable-CNNs, which can be equivariant over an arbitrary number of discrete rotations by using circular harmonics as the basis functions for the equivariant kernels.

Group CNNs assign a c-dimensional feature vector $f(x)$ at each location x in the image plane $\mathbb{R}^2$. A collection of these feature vectors forms a new space $\mathbb{R}^{c_{in} \times c_{out}}$ called a feature field for each position where c_{in} is the number of input channels and c_{out} the number of output channels. In the case of grey-scale images like the ones used in radio astronomy, both input and output channels are scalar (1-channel) and the field is denoted $s : \mathbb{R}^2 \to \mathbb{R}$.

For equivariance to hold, if we transform an input x to a layer Φ in the model, this must be equivalent to transforming the output from the layer

$$\Phi(T_g x) = T'_g \Phi(x) \tag{15}$$

where T_g is the transformation matrix of g and T'_g is not necessarily the same as T_g. We require that T_g is a linear operator such that a set of T_g for a feature field f is called a group representation.

The representation $\rho(g)$ of a transformation g describes how the various channels of each feature vector mix when transformed. For some set of transformations g from a group G, we can say that a given convolutional kernel κ is equivariant if

$$\kappa(gx) = \rho_{out}(g)\kappa(x)\rho_{in}(g^{-1}) \tag{16}$$

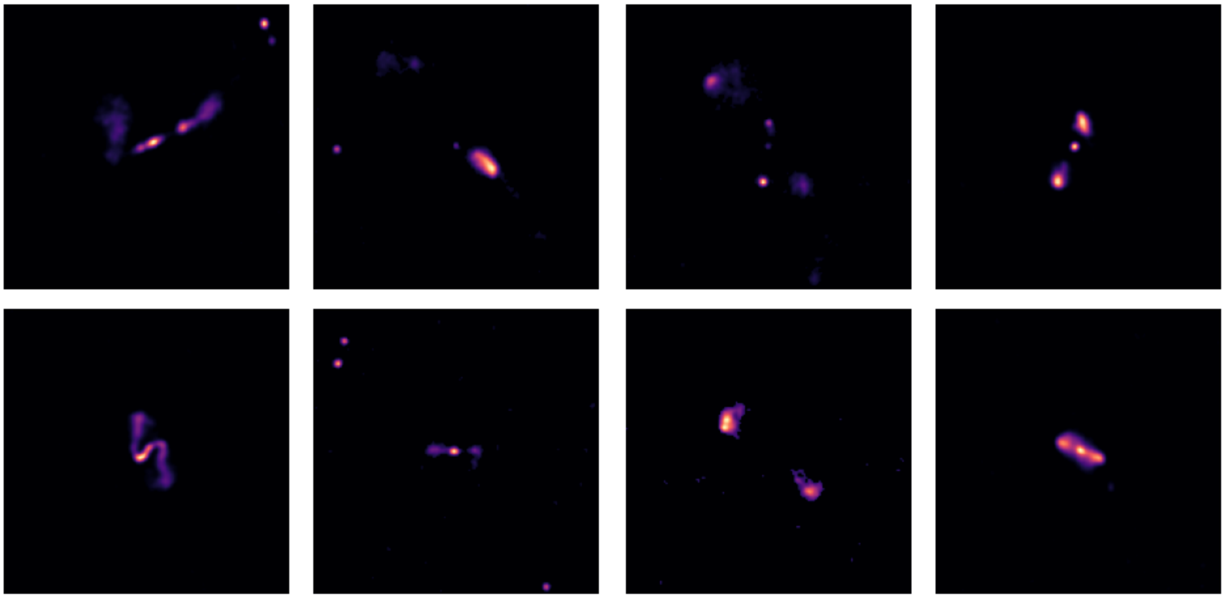
which must hold for all $g \in G$ and all points x in the input plane $\mathbb{R}^2$. In the case of scalar feature fields, $\rho_{in} = 1$ and is called the trivial representation. The output group representation ρ_{out} is defined externally. Kernels that satisfy the constraint in Equation 16 are called steerable and are constructed from a linear combination of basis functions, typically circular harmonics in the case of 2D images.

We take the group G to be the E(2) Euclidean group, which consists of all rotations, reflections and translations in the image plane. Specifically we consider the cyclic, C_N, and dihedral, D_N, subgroups which consist of N discrete rotations of $\frac{2\pi}{N}$ around the image centre. The D_N group also contains reflections about $x = 0$ at each rotation and thus has size $|G| = 2N$. For these selected groups, we use the regular representation ρ_{reg} of the finite group G which simply permutes the channels (axes) of the vector space $\mathbb{R}^{|G|}$, i.e. permutes through the different rotations and reflections. The representation of a transformation group must also commute with any non-linear functions applied after a group convolution. In the case of C_N and D_N symmetry groups, Weiler and Cesa 2019 also prove that these groups commute with all point-wise operations such as ReLU, MaxPool and Average Pooling.

3. Data

Radio galaxies are characterised by an active galactic nucleus (AGN) that results in jets of charged particles that extend out of the centre of a galaxy. [Fanaroff and Riley 1974] proposed a binary classification of these radio galaxies based on the extent of their jets. The separation of the 2 brightest points on either of the galaxy was compared against the total size of the source. If this ratio was below 0.5, the galaxy was classified as FR I, while greater than 0.5 indicated an FR II type galaxy. Figure 1 shows some examples used in our machine learning dataset. Typically FR I galaxies show a bright central lobe, while FR IIs commonly have bright termination shocks at either end of the source.

Models were trained using the MiraBest dataset [Miraghaei and Best 2017], which contains a total of 1329 images of radio sources cross-referenced from the Sloan Digital SKy Survey (SDSS) [Abazajian et al. 2009] with the Faint Images of the Radio Sky at Twenty centimetres (FIRST) [Becker, White, and Helfand 1995] and NRAO VLA Sky Survey (NVSS) [Condon et al. 1998] catalogs. Only sources that were determined to have an AGN were selected as opposed to radio emission from star formation.



(a) FR I galaxies

(b) FR II galaxies

Figure 1: 4 examples of each type of galaxy from the MiraBest Confident test set

Within the FR classification framework different morphologies also exist, as well as unclassifiable objects that have no clear distinctions. Since these classifications were judged by humans, a certain degree of subjectivity was introduced on top of the limited resolution of FIRST data, thus another label was introduced to classify a source as Confident or Uncertain. Of these 1329 sources, 40 were unclassifiable, and 35 found to be hybrid sources with either side of the source falling into different FR types. Additional morphologies were also identified, and these receive their own classifications. Double-double sources consist of 2 closely aligned FR II like jets, which may arise from jet activity restarting after a period of inactivity [Kaiser, Schoenmakers, and Röttgering 2000]. Wide-angle tail sources have their jets swept back into a symmetric 'C' shape by various processes [Sakelliou and Merrifield 2000]. Diffuse sources are also identified. Each image is labelled with 3 numbers that represent different characteristics of the galaxy as shown in Table 1.

For the machine learning dataset, 73 sources were removed, including the unclassified sources and all sources that were larger than the required 150 × 150 image size. The dataset was pre-processed as outlined in Aniyan and Thorat 2017. This involved centering the image on the source and cropping to 150 × 150 pixels. This was then followed by noise reduction, which set any pixels to 0 that were below 3 times the local rms noise level. Also, all pixels were set to 0 outside of a circular mask of radius 75 pixels centred on the source to allow for data augmentation in the form of rotations to be performed. The images were finally normalised using the difference between the minimum and maximum pixel values such that the brightest pixel had a value of 255 and the darkest a value of 0.

The MiraBest Confident dataset is a subset of the data with all sources labelled as uncertain and hybrid removed. There is no distinction made between sub-classes such that only FR I and FR II sources remain. The data is divided into training and test data, the training data is then split 80:20 into a training set and validation set respectively. The

Digit 1	**Digit 2**	**Digit 3**
1: FR I	0: Confident	0: Standard
2: FR II	1: Uncertain	1: Double-double
3: Hybrid		2: Wide-angle tail
4: Unclassifiable		3: Diffuse
		4: Head-tail

Table 1: 3 Digit classification scheme for images in the MiraBest dataset

MiraBest Confident test set contains a total of 104 images, with 49 classified as FR I and 55 FR II sources with a similar proportion in the training data. The MiraBest Uncertain dataset may also be used to test the trained model on out-of-distribution images, and it contains 49 galaxies labelled as uncertain. 25 of these are FR I while the remaining 24 are FR II.

4. Models

In this work we use a modified LeNet-5 [Lecun et al. 1998] architecture with 2 convolutional layers as our conventional CNN. For our equivariant CNNs, the standard convolutional layers are replaced by steerable convolutional layers from the e2cnn extension library [1]. We use rotational orders of 4, 8 and 16 for the steerable kernels. The convolutional layers are followed by 3 fully-connected layers with MC-dropout ($p_d = 0.5$) enabled on the final layer. Since we use MC-dropout, we expect over-fitting to be significantly reduced during training [Srivastava et al. 2014]. Table 2 shows a brief outline of the layers used and the number of channels.

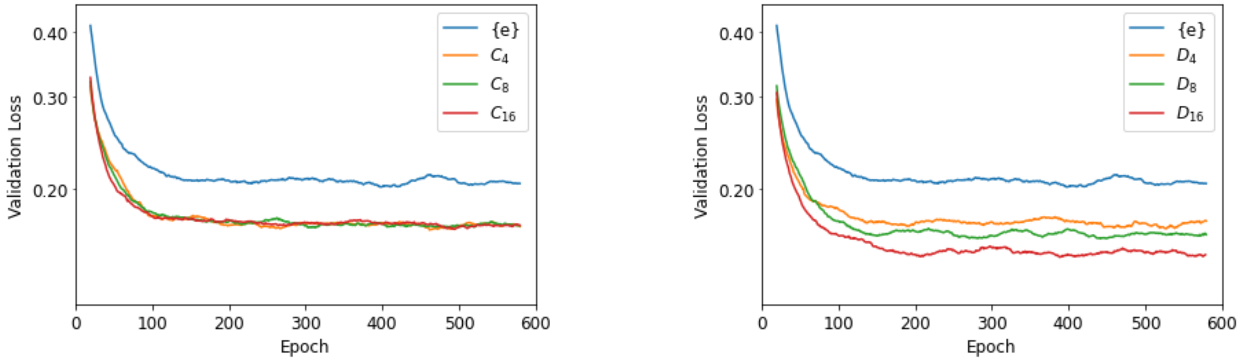


Figure 2: Validation loss plotted over 600 training epochs for LeNet model {e} and all Cyclic-equivariant C_N (left) and Dihedral-equivariant D_N (right) models where N denotes the order of equivariance.

Our models were trained over 600 epochs on the MiraBest Confident dataset, using a batch size of 50, an initial learning rate of 10^{-4}, and a weight decay value of 10^{-6}. Early stopping is enabled during training, the model with the lowest validation error is used. Figure 2 shows the validation loss of each model during training. It can be seen that there

[1] https://github.com/QUVA-Lab/e2cnn

Operation	**Kernel**	**Channels**
Invariant Projection		
Convolution	5×5	6
ReLU		
MaxPool	2×2	
Convolution	5×5	16
ReLU		
MaxPool	2×2	
Invariant Projection		
Average Pool		
Fully-connected		120
ReLU		
Fully-connected		84
ReLU		
MC-dropout		
Fully-connected		2
(Spectral Normalisation)		

Table 2: Brief overview of model architecture used. Any operations in italics are only active in the equivariant models. All convolutional layers have a padding of 1 empty pixel added to the sides of the images and stride of 1 pixel. Spectral normalisation is optionally applied to the final fully-connected layer. The final layer output is then passed into a softmax function to provide the final classification probability.

is no improvement in validation loss for cyclic models above order 4, whereas the dihedral models show a considerable decrease in the loss function for increased model orders. There is also quicker convergence to a lower validation loss by both cyclic and dihedral models of higher orders, attributed to the improved ability of these networks to generalise [Weiler and Cesa 2019]. For the second part of our results, spectral normalisation was applied to the final fully-connected layer in order to normalise the weight matrix by the largest singular value. The class-wise and total accuracy of each model on the reserved test set is displayed in Table 3 for both the unmodified models and with spectral normalisation applied. We observed slight increases in FR II accuracy using dihedral models over cyclic models, while FR I performance remained almost the same which was also observed by Scaife and Porter 2021.

Comparisons between the numerous uncalibrated equivariant networks and the standard LeNet model are shown in Table 4, where the number of samples are shown for which the absolute value of confidence/uncertainty has changed by more than 0.01. All models show a net decrease in uncertainty, while almost all models have a net increase in confidence, with the exception of C_8 and C_{16}. The D_4 model has the greatest increase in confidence and decrease in uncertainty, which may be a result of the model becoming overconfident. However, we later show that this is in fact one of the best calibrated models we were able to produce.

	Original			Spectral Normalisation		
Model	FR I	FR II	Total	FR I	FR II	Total
{e}	93.34	94.27	93.90	93.83	95.00	94.45
C_4	94.67	94.06	94.40	**95.79**	94.09	94.89
C_8	**96.06**	93.75	94.88	95.34	95.78	95.57
C_{16}	95.82	94.56	95.19	95.29	95.66	95.49
D_4	95.35	**97.51**	**96.54**	95.51	96.94	96.27
D_8	95.16	94.10	94.60	95.56	**96.98**	**96.36**
D_{16}	95.41	95.79	95.65	95.43	94.29	94.87

Table 3: Class-wise and total accuracy as a function of model order. All values are a % and represent the performance on the reserved test set. Highlighted in bold are the highest accuracy values for a given category.

	Confidence		Uncertainty $\langle\eta\rangle$	
Model	**Increased**	**Decreased**	**Increased**	**Decreased**
C_4	39	11	7	21
C_8	23	34	16	18
C_{16}	26	32	10	16
D_4	45	14	7	22
D_8	36	16	9	20
D_{16}	35	24	8	18

Table 4: Number of samples that saw an increase or decrease of more than 0.01 in absolute confidence/uncertainty when compared to the standard LeNet model.

5. Results

5.1. Initial Results

The ECE score indicates how well the output probability aligns with the true probability of classification. We compute all of our ECE calibration metrics using a total of 13 bins, selected such that the 104 test image are placed into equal-size bins of magnitude 8. Figure 3 displays example ECE plots for 3 models, with all plots following a similar shape that resembles a sigmoid function, indicating a small amount of under-confidence in every model. ECE scores for the remaining models are shown in Table 5. The best calibrated model was found to be the D_{16} model, with an ECE of 2.37%.

We quantify the uncertainty in each prediction using both the overlap index and predictive entropy in order to compare between the two metrics and how they represent the uncertainty. From observing the produced UCE plots in Figure 4 and using predictive entropy as our uncertainty, we find that all models consistently provide high uncertainties on images with few to no misclassifications, showing that this metric causes massive under-confidence. However, when using the overlap index as a measure of uncertainty the UCE scores were much lower, and all models appear reasonably well calibrated when using this metric. There is a slight overconfidence in classifying high uncertainty sources when using the overlap index that can be seen in the form a slight deviation from the dashed line on the top row of Figure 4 and is present for all models.

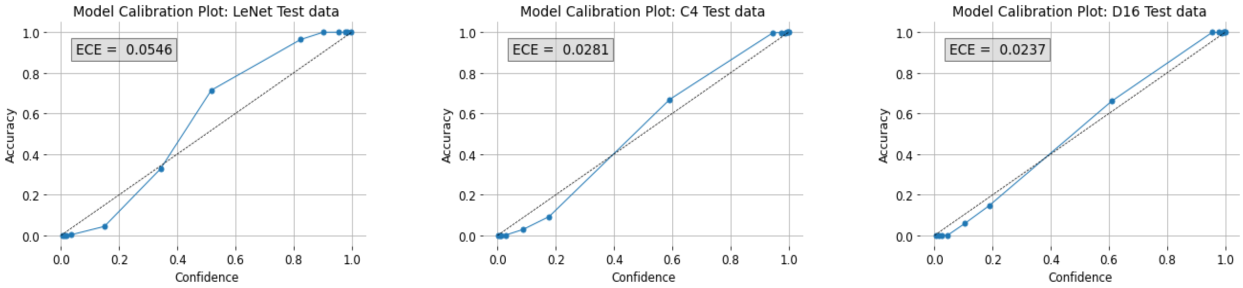


Figure 3: ECE plots for 3 different models. Confidence is the softmax probability of an FR II classification while accuracy is the proportion of such classifications. All plots are made with M=13 bins of size $|B_M| = 8$. A perfectly calibrated model is represented by the dashed line, implying no difference between confidence and accuracy.

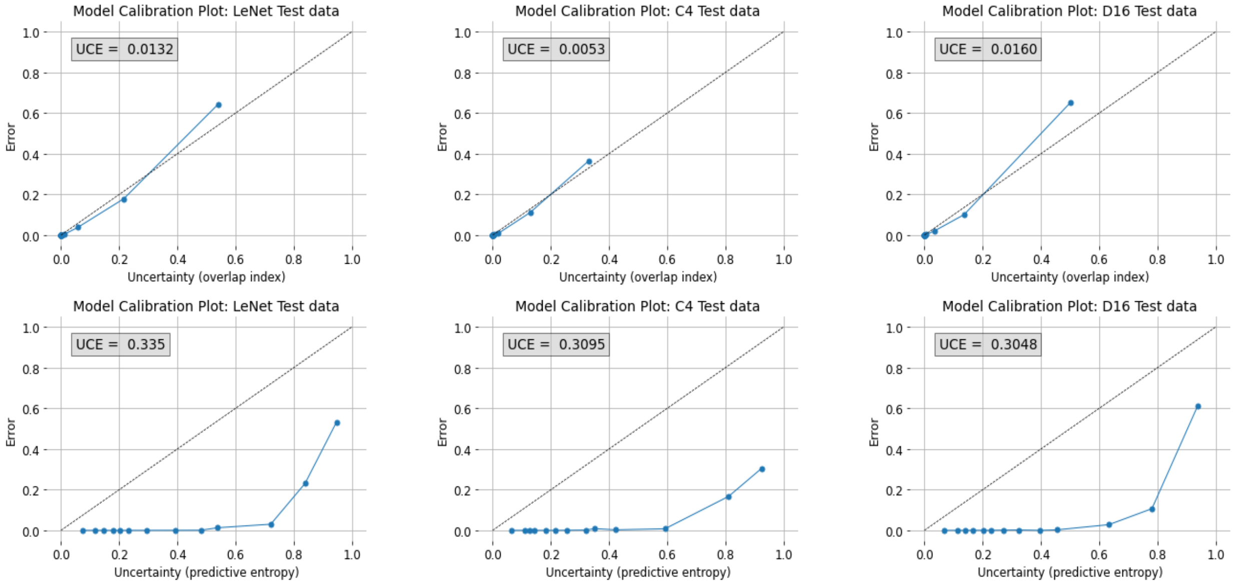


Figure 4: (top) UCE plots for LeNet, C_4 and D_{16} models using the index free overlap value as the uncertainty metric. (bottom) UCE plots for the same models using predictive entropy as uncertainty. $M = 13$ bins of equal size $|B_M| = 8$ are used for all plots.

It is possible to observe samples with very low overlap but high predictive entropy, implying that the entropy is not a useful metric to encapsulate the number of mis-classifications without calibration applied. This is highlighted in Figure 5 where we display the distribution of softmax probabilities across 9 rotation augmentations. The overlap index is much closer to the actual error rate than the entropy. For samples with high error rates, it can be unclear as to which label the model is trying to assign as shown on the right in Figure 5.

5.2. Model Adaptations - Spectral Normalisation

Spectral Normalisation is one technique used extensively in Generative Adversarial Networks in order to stabilise the training process and increase model robustness against adversarial attacks, which involve adding noise to the test image in order to change the prediction [Miyato et al. 2018]. Mukhoti et al. 2021 demonstrated that the introduction of spectral normalisation can lead to better identification of out-of-distribution (OoD) data. It has also

	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: values of ECE, UCE using overlap index and UCE using entropy for unmodified models (left) and models with spectral normalisation applied on the final layer (right). The best calibration score is highlighted in bold.

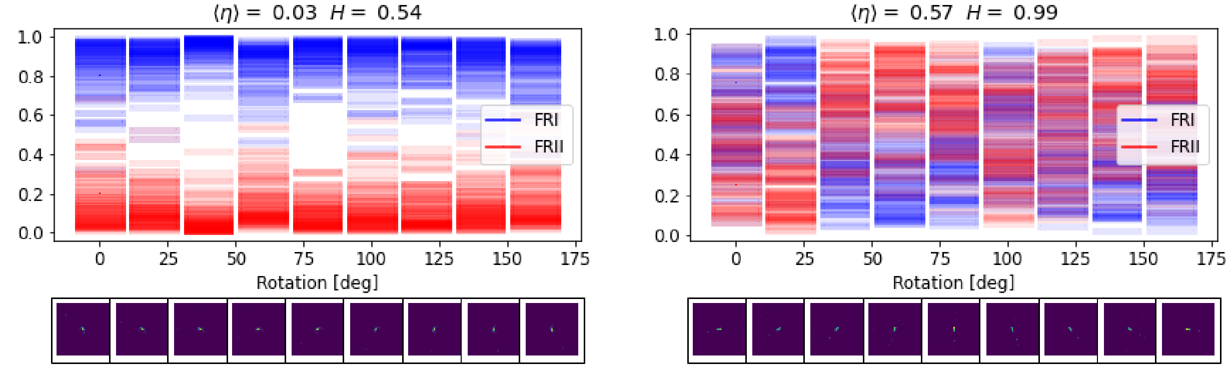


Figure 5: Example softmax probability distributions from the D_4 model across 9 rotations of 20 degrees shown on the bottom. The overlap index and predictive entropy of each distribution are displayed above. (left) Average error rate: 1%. (right) Average error rate: 40.3%.

been suggested in Scaife and Porter 2021 that certain equivariant orders perform slightly better than other models due to discretisation artifacts in the small convolutional kernels of size 5 × 5. Spectral normalisation may alleviate this as it reduces the networks sensitivity to small changes such as these.

The Lipschitz constant of a continuously differentiable function represents the maximum absolute gradient connecting any 2 neighbouring points in the graph of the function. Normalising the function by the Lipschitz constant thus decreases the sensitivity to input perturbations. For a linear layer of a neural network $g(h) = \mathbf{W}h$ acting on some input h where $\mathbf{W}$ is the weight matrix of the layer, the Lipschitz normalisation constant $||g||_{Lip}$ is equal to the largest singular value $\sigma_1(\mathbf{W})$ of the matrix [Miyato et al. 2018]. To obtain $\sigma_1(\mathbf{W})$, we use singular value decomposition (SVD) which can be applied to any $m \times n$ matrix. $\mathbf{W}$ can thus be factorised as follows

$$\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{17}$$

where $\mathbf{U}$ is the $m \times m$ unitary matrix of left-singular vectors, $\mathbf{V}^T$ the transposed $n \times n$ unitary matrix of right-singular vectors and $\mathbf{\Sigma}$ is an $m \times n$ diagonal rectangular singular value matrix.

The largest singular value $\mathbf{\Sigma}_{11} = \sigma_1$ is used to normalise the elements of the weight matrix $\mathbf{W}$ and thus increase the 'Lipschitzness' or smoothness of the weight matrix with regard to any inputs. We use the built-in spectral normalisation function from the Pytorch library and use $n = 1$ power iterations to quickly approximate the largest singular value [Pan and Mishra 2021]. Spectral Normalisation was only applied to the final fully-connected layer of our models and each model was then retrained with this implemented.

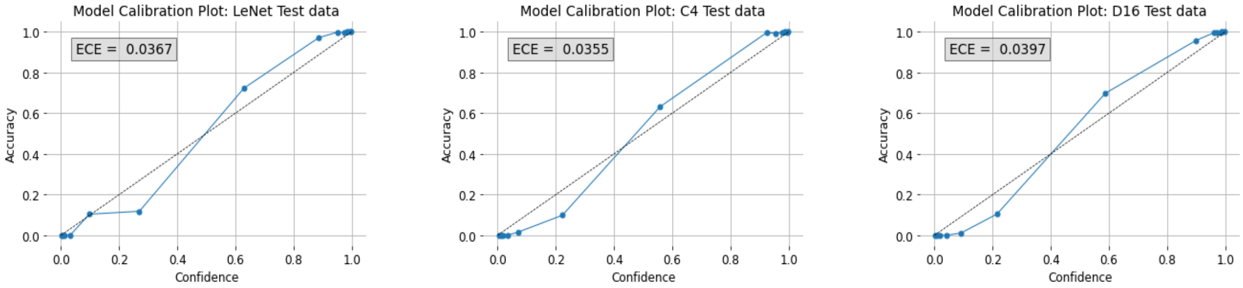


Figure 6: ECE plots with spectral normalisation applied for 3 different models. All plots are made with M=13 bins of size $|B_M| = 8$.

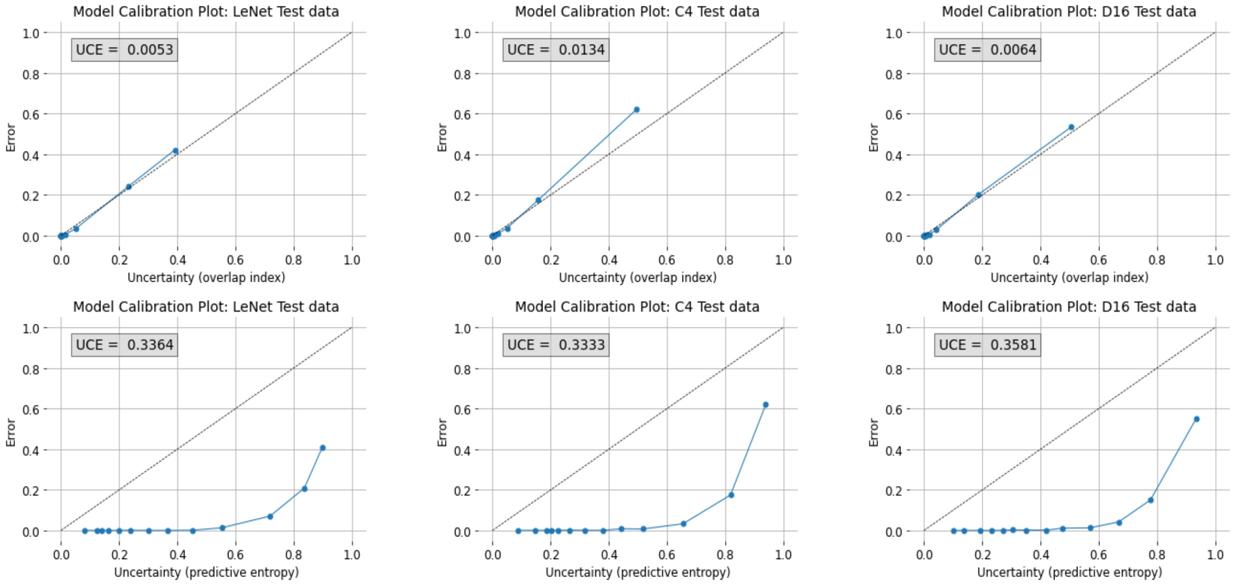


Figure 7: (top) UCE plots for spectrally normalised LeNet, C_4 and D_{16} models using the index free overlap value as the uncertainty metric. (bottom) UCE plots for the same models using predictive entropy as uncertainty. $M = 13$ bins of equal size $|B_M| = 8$ are used for all plots.

Figure 6 shows the ECE calibration curves for the same three example models used before with spectral normalisation enabled on the final fully connected layers of each model. Again, a noticeable sigmoid shape appears which indicates under-confidence and is present for all model orders. Figure 7 displays the corresponding UCE plots with spectral normalisation applied. All graphs follow similar forms to those without spectral normalisation applied.

6. Discussion

6.1. Analysis of Calibration Metrics

It has been shown by Guo et al. 2017 that large deep learning models often suffer from poorly calibrated outputs, thus it was expected that the models used here would provide reasonably well-calibrated probabilities as the architectures used are much smaller (5 layers) compared to more modern neural networks such as ResNet-100. Niculescu-Mizil and Caruana 2005 were also able to show that binary classification neural networks are naturally inclined to be well calibrated. This is supported by the low ECE scores observed on the order of 3% while also seeing the characteristic sigmoid shape in the plots. We find that all of the equivariant models are better calibrated than the classic LeNet architecture which had a probability calibration error of 5.46%, however it is important to note that there is no clear trend in the calibration scores as a function of equivariant model order.

Using the overlap index to measure the uncertainty in the model output provides well calibrated results with a maximum calibration error of only 1.6% while the lowest error was 0.53%. This indicates that the overlap is already well suited to describing the model uncertainty without any calibration techniques being applied. The overlap appears far superior to the entropy as a measure of uncertainty as the entropy UCE is consistently around 30% for all models. This may arise due to the overlap assuming a Gaussian distribution of MC-dropout forward passes, whereas the predictive entropy does not consider any distribution, and is limited to just the average model confidence. This can be seen in the left image of Figure 5, where a predictive entropy of 0.54 is calculated due to a low average confidence rather than high spread of data. Another reason that the overlap may be well-calibrated is that it uses a Gaussian kernel to approximate the overlapping in a Gaussian mixture model. If we were to change our prior distribution to, for example, a Laplace prior, it remains to be seen whether the overlap index would still encapsulate the uncertainty using the same kernel.

We performed an additional 2 repeats with the same model to determine the standard error on the ECE and UCE with respect to the use of stochastic dropout. This was found to be on the order of $\approx 1\%$ for all models, suggesting $N = 100$ forward passes were sufficient to provide a precise value of these metrics for a given model. We then trained 2 new LeNet models with different initial weights and found the standard deviation in the ECE to be on the order of $\approx 6\%$. The UCE $\langle \eta \rangle$ was found to be $1.33 \pm 0.23\%$. Although we found an error of 17% on our overlap UCE values, this may simply be due to the fact that the UCE is naturally very small, on the order of 1%. We assume that this error will also be of a similar magnitude for other models, thus making it difficult to comment on any observed trends within our data that uses just one trained example of each model. It is suggested that this high variance may come from the inherent issues with MC-dropout as a Bayesian approximation, outlined in section 6.4.

Another issue with using the ECE to consider calibration is the high sensitivity to the number of bins M used in the histogram as shown by Laves et al. 2021, as the ranking of models is therefore also sensitive to M. Conversely, the UCE has no such issue and is entirely independent of M, meaning any comparisons drawn between models using this metric are more represent of the true calibration of the models.

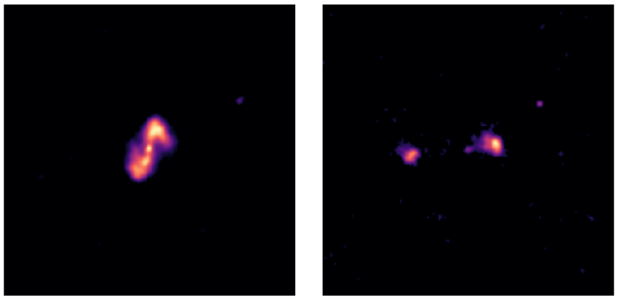


Figure 8: (left) an FR I type galaxy that is consistently mis-classified at a rate of $> 90\%$ by all models as FR II with very low uncertainty. (right) an FR II labelled galaxy that is often classified as FR I with $\sim 80\%$ confidence by most models.

The slight under-confidence in the models as shown by the ECE plots may be fixed by introducing calibration such as applying temperature scaling to the confidence [Guo et al. 2017] or implementing Platt scaling to each range of probabilities [J. C. Platt 1999]. It can also be noted that for predictive entropy values ≤ 0.5, the error rates very rarely rose above 0% and so a form of scaling should also be implemented for the entropy to calibrate it.

One possible addition involves implementing a threshold uncertainty/confidence level for histogram binning to provide more comparable results between models by only considering samples above this threshold in the calculation of the ECE. This would remove any advantage given to more accurate (or over-confident) models, which naturally have more binned points at the extreme ends of the graphs, with fewer uncertain points contributing to the error.

6.2. Analysis of Model Adaptations

No significant differences in calibration scores were observed in the spectral normalised models compared to their unmodified counterparts and likewise, no obvious trends were observed across model orders. This is likely due to only using one instance of each model for our tests and so our results are subject to variation in weight initialisation and data augmentations during training, resulting in slightly differing models. We plan to alleviate this in further experiments by using multiple trained versions of each architecture. We did observe that the minimum ECE and entropy UCE shown in Table 5 were lower when using spectral normalisation for the C_8 model, suggesting spectral normalisation may allow for the capacity for better calibrated models to be trained.

Overall, our findings indicate that spectral normalisation generally has no significant effect on the calibration of these neural networks. There has been recent experiments with convolutional normalisation in which the kernels themselves are normalised in a similar way [Liu et al. 2021]. This further improves the Lipschitzness of the network and may be one reason why we were not able to see a significant difference with our normalised models as our form of normalisation is applied only to the final fully-connected layer.

6.3. Notes on data selection

Figure 8 displays two examples of potential labelling error in the dataset. The FR I galaxy is consistently mis-classified by all models with very high confidence which suggests it may be a case of label noise in the dataset. The FR II galaxy may be mis-classified due to over-fitting of the model parameters as some models were able to correctly identify it as FR II but with low confidence.

Since we used the MiraBest Confident dataset, most samples are confidently labelled correctly hence we observe a lack of high uncertainty data. We cannot extrapolate the calibration curves of these models at low uncertainty to high uncertainty or OoD data and so further experiments with more uncertain data-sets is needed to confirm that the overlap index is indeed well calibrated. The predictive entropy can be broken down into aleatoric and epistemic uncertainty, which could be used for OoD detection by identifying sources with higher than usual aleatoric uncertainty. We also note that multiple samples from the Confident dataset were assigned a predictive entropy close to 1, which suggests maximum uncertainty for these images, yet matching error rates were not observed. It has also been shown that spectral normalisation leads to robust OoD detection for image classification tasks on CIFAR-100 by Mukhoti et al. 2021, which warrants further investigation with our models using different test sets to see if this also holds for Fanaroff-Riley classification.

6.4. Potential issues with MC-dropout

It has been argued that the use of MC-dropout to approximate uncertainties does not actually represent the underlying Bayesian model uncertainties [Folgoc et al. 2021]. Instead, MC-dropout produces a multi-modal posterior that has no correlation to the uncertainties in the true posterior distribution. Folgoc et al. 2021 show that for a simple 1-layer regression task the Kullbach-Leibler divergence, which quantifies how well one distribution approximates another, approaches $+\infty$ for MC-dropout. Instead it is proposed that variational inference should be used to approximate the true posterior and its uncertainties while being much more costly to compute. Our findings show that using MC-dropout to calculate an overlap index provides a well calibrated way of representing the error rate as an uncertainty while also remaining relatively cheap to compute, in direct contrast these issues. However, the use of MC-dropout may be resulting in different approximations of the Bayesian uncertainties every time a model is trained, thus causing a large variance in our calibration scores for similar models.

7. Conclusions

Using MC-dropout to approximate an underlying Bayesian model, we have confirmed previous findings that show equivariance provides better accuracy, and improved FR II classification accuracy for dihedral-equivariant models. We have also shown that the implementation of group equivariance leads to better confidence calibration over non-equivariant models when classifying Fanaroff-Riley type radio galaxies. It was observed that all equivariant models, regardless of order, were similarly well-calibrated using the ECE metric. However the inherent issues with using the ECE lead us to look into the more robust UCE.

The uncertainty calibration of our models all indicate very well calibrated outputs regardless of equivariance when using the overlap index as an uncertainty measure. The overlap index therefore provides a reliable way of quantifying the expected error of a sample, allowing for identification of potentially mis-classified and uncertain sources when used on unlabelled data. The MiraBest data-set that was used has been heavily curated and only confidently labelled sources were included during training and test time, thus our models were not exposed to the true distribution of radio galaxies. It is imperative that more tests be performed with data-sets that contain more uncertain or out-of-distribution samples as we observed a slight under-estimation of uncertainty for the most ambiguous data. Two sources of such data include the MiraBest Uncertain and MiraBest Hybrid test sets.

When using the entropy as an uncertainty as is common in machine learning, we observed poorly calibrated results that do not reflect the actual error rate without some form of calibration applied. We attribute this to the overlap index being able to approximate the actual distribution of MC samples, rather than simply considering the distance between the mean values of the softmax probability distributions.

We performed the first application of spectral normalisation to FR classification. Our models showed no improvement in calibration significant enough to comment on in the tests that we were able to perform. With more time it would be possible to take averages of an ensemble of models with the same architecture to examine this further. It is also hoped that spectral normalisation will allow for more robust OoD data detection.

We have applied the individual error rates in conjunction with our measures of uncertainty to identify a potential case of label noise in the data-set that we used, while also identifying a mis-classified but uncertain source. This highlights the usefulness of uncertainty and confidence metrics in deep learning as ways to analyse the data-sets being used and improve them to lead to better model development.

References

Abazajian, Kevork N. et al. (June 2009). “The Seventh Data Release of the Sloan Digital Sky Survey”. In: 182.2, pp. 543–558.

Aniyan, A. K. and K. Thorat (June 2017). “Classifying Radio Galaxies with the Convolutional Neural Network”. In: *The Astrophysical Journal* 230.2, 20, p. 20.

Becker, Robert H., Richard L. White, and David J. Helfand (Sept. 1995). “The FIRST Survey: Faint Images of the Radio Sky at Twenty Centimeters”. In: *The Astrophysical Journal* 450, p. 559.

Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe (Apr. 2017). “Variational Inference: A Review for Statisticians”. In: *Journal of the American Statistical Association* 112.518, pp. 859–877.

Cohen, Taco S. and Max Welling (2016). *Steerable CNNs.*

Condon, J. J. et al. (May 1998). “The NRAO VLA Sky Survey”. In: *The Astrophysical Journal* 115.5, pp. 1693–1716.

Dieleman, Sander, Jeffrey De Fauw, and Koray Kavukcuoglu (2016). *Exploiting Cyclic Symmetry in Convolutional Neural Networks.*

Fanaroff, B. L. and J. M. Riley (May 1974). “The morphology of extragalactic radio sources of high and low luminosity”. In: *Monthly Notices of the Royal Astronomical Society* 167, 31P–36P.

Folgoc, Loic Le et al. (2021). *Is MC Dropout Bayesian?*

Gal, Yarin (2016). *Exploring Uncertainty Measures in Deep Networks for Multiple Sclerosis Lesion Detection and Segmentation.*

Gal, Yarin and Zoubin Ghahramani (2015). *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning.*

Guo, Chuan et al. (2017). *On Calibration of Modern Neural Networks.*

Izmailov, Pavel et al. (18–24 Jul 2021). “What Are Bayesian Neural Network Posteriors Really Like?” In: *Proceedings of the 38th International Conference on Machine Learning.* Proceedings of Machine Learning Research 139. Ed. by Marina Meila and Tong Zhang, pp. 4629–4640.

Jarvis, Matt J. et al. (May 2016). “The MeerKAT International GHz Tiered Extragalactic Exploration (MIGHTEE) Survey”. In: *MeerKAT Science.* Vol. MeerKAT2016. Stellenbosch, South Africa, p. 006.

Johnston, S. et al. (Dec. 2008). “Science with ASKAP. The Australian square-kilometre-array pathfinder”. In: *Experimental Astronomy* 22.3, pp. 151–273.

Kaiser, Christian R., Arno P. Schoenmakers, and Huub J. A. Röttgering (June 2000). “Radio galaxies with a ‘double-double’ morphology—II. The evolution of double-double radio galaxies and implications for the alignment effect in FRII sources”. In: *Monthly Notices of the Royal Astronomical Society* 315.2, pp. 381–394.

Khan, Salman, Munawar Hayat, and Fatih Porikli (2017). *Regularization of Deep Neural Networks with Spectral Dropout.*

Kristiadi, Agustinus, Matthias Hein, and Philipp Hennig (2020). *Being Bayesian, Even Just a Bit, Fixes Overconfidence in ReLU Networks.*

Laves, Max-Heinrich et al. (2019). *Well-calibrated Model Uncertainty with Temperature Scaling for Dropout Variational Inference.*

— (2021). *Uncertainty Calibration Error: A New Metric for Multi-Class Classification.*

Lecun, Y. et al. (1998). “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11, pp. 2278–2324.

Liu, Sheng et al. (2021). “Convolutional Normalization: Improving Deep Convolutional Network Robustness and Training”. In: *Advances in Neural Information Processing Systems* 34. Ed. by M. Ranzato et al., pp. 28919–28928.

McConnell, D. et al. (2020). “The Rapid ASKAP Continuum Survey I: Design and first results”. In: *Publications of the Astronomical Society of Australia* 37, e048.

Miraghaei, H. and P. N. Best (Jan. 2017). “The nuclear properties and extended morphologies of powerful radio galaxies: the roles of host galaxy and environment”. In: *Monthly Notices of the Royal Astronomical Society* 466.4, pp. 4346–4363.

Miyato, Takeru et al. (2018). *Spectral Normalization for Generative Adversarial Networks.*

Mohan, Devina et al. (Jan. 2022). “Quantifying uncertainty in deep learning approaches to radio galaxy classification”. In: *Monthly Notices of the Royal Astronomical Society* 511.3, pp. 3722–3740.

Mukhoti, Jishnu et al. (2021). *Deep Deterministic Uncertainty: A Simple Baseline.*

Namdari, Alireza and Zhaojun (Steven) Li (2019). “A review of entropy measures for uncertainty quantification of stochastic processes”. In: *Advances in Mechanical Engineering* 11.6, p. 1687814019857350.

Niculescu-Mizil, Alexandru and Rich Caruana (2005). “Predicting Good Probabilities with Supervised Learning”. In: *Proceedings of the 22nd International Conference on Machine Learning.* ICML ’05. Bonn, Germany: Association for Computing Machinery, pp. 625–632.

Nixon, Jeremy et al. (2019). *Measuring Calibration in Deep Learning.*

Noci, Lorenzo et al. (2021). “Disentangling the Roles of Curation, Data-Augmentation and the Prior in the Cold Posterior Effect”. In.

Ntwaetsile, Kushatha and James E Geach (Feb. 2021). “Rapid sorting of radio galaxy morphology using Haralick features”. In: *Monthly Notices of the Royal Astronomical Society* 502.3, pp. 3417–3425.

Pan, Zhixin and Prabhat Mishra (2021). *Fast Approximate Spectral Normalization for Robust Deep Neural Networks.*

Pastore, Massimiliano and Antonio Calcagnì (2019). “Measuring Distribution Similarities Between Samples: A Distribution-Free Overlapping Index”. In: *Frontiers in Psychology* 10.

Platt, John C. (1999). “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods.” In: *Advances in Large Margin Classifiers* 10(3), pp. 61–74.

Sakelliou, Irini and Michael R. Merrifield (Jan. 2000). “The origin of wide-angle tailed radio galaxies”. In: *Monthly Notices of the Royal Astronomical Society* 311.3, pp. 649–656.

Scaife, Anna M. M. and Fiona Porter (Feb. 2021). “Fanaroff–Riley classification of radio galaxies using group-equivariant convolutional neural networks”. In: *Monthly Notices of the Royal Astronomical Society* 503.2, pp. 2369–2379.

Simard, Patrice Y., Dave Steinkraus, and John Platt (Aug. 2003). “Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis”. In.

Srivastava, Nitish et al. (2014). “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”. In: *Journal of Machine Learning Research* 15.56, pp. 1929–1958.

Weiler, Maurice and Gabriele Cesa (2019). “General E(2)-Equivariant Steerable CNNs”. In: *CoRR* abs/1911.08251.

Weiler, Maurice, Fred A. Hamprecht, and Martin Storath (2018). “Learning Steerable Filters for Rotation Equivariant CNNs”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 849–858.