Uncertainty Quantification and Calibration of Deep Learning Models for Astrophysical Analysis

Samuel Hudson 10458093

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

Abstract

The influx of data from new radio telescopes requires efficient and accurate labelling which can be fulfilled using machine learning, however, designing robust and well-calibrated learning models is critical for scientific applications. Radio galaxy morphology is typically separated using the binary Fanaroff-Riley categorisation scheme and we apply group-equivariant CNNs to classify the galaxies in the MiraBest dataset. Recent studies have revealed considerable overlap in the properties of the galaxies; quantifying this ambiguity via uncertainty measures is essential to understanding the classification scheme. In this work we explore different uncertainty methods for deep neural networks, initially utilising MC-dropout to approximate a Bayesian posterior. We then implement a Gaussian Process classifier in conjunction with Batch Normalisation of the convolutional layers to achieve 99.1% accuracy on a reserved test set. The calibration of uncertainties is compared against that of MC-dropout architectures using the ECE and UCE metrics. Temperature scaling is shown to calibrate deep learning models extremely well and we find that using the overlap index to quantify the uncertainty provides a calibration error $< 0.5\%$. We use the models to successfully isolate out of distribution images of optical galaxies by assigning low confidence scores and finally study the relationship between physical properties (radio luminosity and size) and the resulting morphology of FR type radio galaxies.

Contents

1. Introduction

A class of new radio telescopes recently came online (ASKAP [McConnell et al. 2020], MeerKAT [Jarvis et al. 2016]) as precursors to the Square Kilometer Array. The output is vast amounts of data that cannot feasibly be studied by hand. Instead, automated algorithms have been employed to assist in processing this data while highlighting any interesting samples for closer inspection. Bayesian deep learning methods fill the requirement for fast and reliable analysis of this data, however naive application can lead to incorrect predictions and important features being overlooked. Uncertainty calibration attempts to fix this by tuning the models in a test environment to ensure that the probabilities and uncertainties assigned by the models are meaningful, also allowing any sources of interest to be highlighted by the model.

1.1. Radio Astronomy

The most commonly used classification scheme for radio spectrum galaxies was first defined by Fanaroff and Riley 1974, in which galaxies are placed into 2 categories based on the extent of their emission in relation to their size. FRI type galaxies are brighter towards the galactic centre, with the separation between the brightest points less than half of the full extent of the source. FRII sources have more intense emission further from the nucleus and their brightest regions are spaced by more than half of the full size of the source. Fanaroff and Riley 1974 originally found that FRII galaxies are generally brighter than FRI sources, with a defining boundary of 10^{25} WHz^{-1} at 178 MHz. Further studies with more sensitive radio telescopes such as LOFAR (150MHz) [Mingo et al. 2019] have enabled detection of fainter and lower surface-brightness FR sources which has unveiled a large overlap in the size and luminosities of these objects. In particular, the discovery of a significant population of low luminosity FR II galaxies has challenged the view that these 2 classes can be so easily linearly separated.

The physical differences in FRI and FRII type galaxies are not well understood, but it has been suggested that they can be attributed to the surrounding galactic environment, with denser, richer gas leading to disruption of the relativistic jets closer to the Active Galactic Nucleus (AGN) in the case of FRI galaxies [Kaiser and Philip N. Best 2007]. FRII galaxies on the other hand would experience very little interference and the jets instead terminate far from the host galaxy, resulting in distinctly separated bright lobes in the radio spectrum. Other possibilities have been discussed in literature, notably by [Baum, Zirbel, and O'Dea 1995] which discusses the morphological variety in terms of the actual properties of the AGN rather than the environment. Ledlow and Owen 1996

discovered a correlation between the optical and radio luminosities of FR type galaxies and was an important step in relating the physical properties of the host galaxies to their radio spectrum morphology. They discuss a bi-variate luminosity function, in which the class division can be defined as a function of the optical luminosity ($\propto L_{opt}^2$), providing further support for the hypothesis that the galactic properties determine the overall appearance of these galaxies.

1.2. Deep Learning

In recent years the capabilities of Deep Learning (DL) models have improved exponentially and they are increasingly implemented into scientific contexts to help analyse huge amounts of data. Despite their powerful predictive and generative capabilities, Deep Learning models can suffer from a multitude of issues, such as over-fitting to the available training data and providing predictions that are under or overly confident [Guo et al. 2017]. Bayesian Neural Networks (BNNs) [Gal and Ghahramani 2016] can solve much of these issues by providing probabilistic predictions with uncertainties. It is preferable for these probabilities and uncertainties to align with the actual chance of mis-identifying data and to represent either noisy images or inherent differences from the training data. Unfortunately these uncertainties are not always well-calibrated and additional tuning must be performed in order to align the model.

In our previous work [Hudson and Millicheap 2023] we discussed and derived the necessary framework for Bayesian learning models. Monte Carlo (MC) dropout was shown by Gal and Ghahramani 2015 to be a fast and efficient way to approximate a Bayesian model by randomly turning off a proportion of the weights in a given layer, thereby simulating an ensemble of models with different parameters. It also has the added benefit of limiting over-fitting and improving generalisation. Simulating multiple models then allows for uncertainty measures to be calculated from their disagreement in predictions which we can then calibrate against the models performance.

One alternative to Neural Networks are Gaussian Processes [Gibbs and Mackay 2000], in which the model learns to fit a multivariate Gaussian probability distribution to the training data through a set of latent functions, allowing unseen data to be classified with both well-calibrated probabilities and meaningful uncertainties related to the distribution of the functions at a given point. The main advantage of using such a model is the introduction of 'distance-awareness' or how closely correlated 2 samples of data are to each other. This is extremely useful for machine learning applications in which new data may be outside of the expected range of the training data, allowing it to be flagged with a high uncertainty and investigated further. Gaussian Processes are typically used to generate a family of functions for regression tasks, but this can be extended to classification tasks too with impressive results [Williams and Barber 1998]. Combining the raw predictive power of Neural Networks and statistical robustness of Gaussian Processes is one area of current research [Liu et al. 2020] that we hope to expand upon.

1.3. Application to Radio Galaxy Classification

Convolutional Neural Networks were a crucial breakthrough in the field of machine vision [Lecun et al. 1998], allowing for fast and efficient extraction of important features from an image through the use of convolutional kernels, and then condensing this information into fewer-dimensions via pooling operations. This can reduce the high-dimensionality of an image into a more manageable amount of data that can then be classified. Due to the nature of the convolution operation, CNNs are naturally invariant to translations of an image. However CNNs can be extended to exploit various other symmetries in the data they are trained on. G-CNNs were first introduced by Cohen and Welling 2016, who extended this equivariance under symmetries to generic groups of transformations allowing for better generalisation capabilities. One group of importance is the

E(2) or 2D Euclidean group, which contains all possible rotations, translations and flips. This is of particular interest in the field of astronomy, for which many astronomical objects have no preferred oriented or chirality.

Scaife and Porter 2021 applied E(2)-CNNs in the context of a Fanaroff-Riley radio galaxy classification task and found improved performance over their non-equivariant counterparts, Mohan et al. 2022 further expanded on this by implementing a fully Bayesian approach to classifying radio galaxies using variational inference. Our previous report [Hudson and Millicheap 2023] studied the calibration of some of these equivariant models and found that they were reasonably well-calibrated without any additional tuning, however we discussed potential calibration methods that we now apply in this report.

Contributions: We achieve an order of magnitude improvement in calibration by applying temperature scaling to our previous work on E(2)-equivariant CNNs, our lowest calibration error is just 0.25% compared to a maximum of 5.58% with no calibration. We implement a Spectrally Normalised Gaussian Process (SNGP) model [Liu et al. 2020] which utilises a Random Fourier Feature (RFF) approximation to a Gaussian Process for the purposes of radio galaxy classification and achieve 99.1% accuracy on a reserved dataset. The capabilities for out of distribution data detection are then studied for the various models. Finally, we consider the work done by Beatriz Mingo et al. 2019 on the differences in FRI and FRII luminosities and physical sizes, and implement our model uncertainties to the MiraBest Confident dataset to examine possible links between the physical properties of the galaxies to the ambiguity of their morphology.

2. Uncertainty Methods in Deep Learning

2.1. Deep Learning Methods

Neural Networks are a powerful tool for fitting complex models to intricate data. Deep Neural Networks (DNNs) consist of multiple layers each defined by learnable weights and biases that are optimised to fit some training data. Given a set of images $\boldsymbol{x} = (x_1, ..., x_N)$ and corresponding labels $\boldsymbol{y} = (y_1, ..., y_N)$ where $y_i = [0, 1]$ for a binary classifier, we can compose a training dataset $\mathcal{D}\{x_i, y_i\}_{i=1}^{N}$ of N image-label pairs. For a batch of training images we can compute the cross-entropy loss function, which is measure of how accurate the predictions are. The optimisation process is then done through back-propagation, in which the gradient with respect to the loss function is computed for every parameter in the model. We minimise the loss function using these gradients to update the weights using a learning rate that we define. Non-linearities are crucial in allowing the model to generalise to the data, and are typically introduced into the model via activation functions placed after every layer, commonly a sigmoid function or Rectified Linear Unit (ReLU). Bayesian models typically use the softmax function which acts on the un-normalised outputs from the final layer (commonly referred to as logits) and yields a probability for each class.

Spectral Normalisation (SN) [Miyato et al. 2018] constrains the largest eigenvalue of a weight matrix and has the effect of reducing the Lipschitz constant of the network (A measure of the smoothness of a function). This means small perturbations in the input do not cause a considerable change in the output and can make a model more robust to noise. This can lead to improved performance and better detection of data that lies outside of the target classification problem. We implemented SN to our models in our previous report finding that it had little to no effect on the model calibration and utilise this technique with all our models in this work.

2.2. Temperature Scaling

It is possible to improve the calibration of a model and mitigate under or over-confidence in the outputs Laves et al. 2019 by implementing a form of logistic regression to the network logits

prior to the application of the softmax function. The softmax function acts on logits z_i for a given image x with predicted label y_i and true label c and yields a probability of classification q_i defined as

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

where C is the total number of classes. Temperature Scaling (TS) controls the strength of this regression through a single parameter β known as the temperature and is can be optimised to calibrate the probabilities.

2.3. Gaussian Processes

A Gaussian Process is a non-parametric Bayesian method of making predictions without any assumptions about the underlying distribution of the data. The posterior distribution is given by a multivariate Gaussian distribution

$$\boldsymbol{y}(\boldsymbol{x}|\mathcal{D}) \sim \mathcal{N}(0, \boldsymbol{K}_{ij}) \tag{2}$$

with mean 0 and covariance matrix $\boldsymbol{K}_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ such that elements of $\boldsymbol{K}$ are computed by applying a kernel k to each pair of data-points in the training data. A common choice of covariance kernel is the Gaussian Radial Basis Function (RBF):

$$k(\boldsymbol{x}, \boldsymbol{x}^*) = \exp\left(-\frac{||\boldsymbol{x} - \boldsymbol{x}^*||}{2\sigma^2}\right) \tag{3}$$

where $k(\boldsymbol{x}, \boldsymbol{x}^*)$ acts on all pairs of training data $\boldsymbol{x}$ and unseen data $\boldsymbol{x}^*$, and σ is the RBF scale parameter which controls the distance between which any two inputs are considered similar.

The predictive mean μ^* and variance σ^{*2} of an unseen data sample x^*, for a model conditioned on training data $\mathcal{D}$ is

$$\begin{aligned} \mu^* &= k(x^*, \boldsymbol{x})\boldsymbol{C}^{-1}\boldsymbol{y}, \\ \sigma^{*2} &= k(x^*, x^*) - \boldsymbol{k}(x^*, \boldsymbol{x})\boldsymbol{C}^{-1}\boldsymbol{k}(\boldsymbol{x}, x^*) \end{aligned}$$

where $\boldsymbol{C}$ is the precision matrix, defined as

$$\boldsymbol{C} = \boldsymbol{K} + \sigma^2_{noise}\boldsymbol{I} \tag{4}$$

with σ^2_{noise} related to the aleatoric uncertainty in the training data [Adlam, Snoek, and Smith 2020]. The predictive posterior label y^* is therefore

$$y^*(x^*|\mathcal{D}) \sim \mathcal{N}(\mu^*, \sigma^{*2}). \tag{5}$$

2.4. Random Fourier Feature Approximation

Since the inversion operation of the matrix $\boldsymbol{C}$ is costly to perform and scales quickly with the number of training samples, we can make an approximation of the kernel and thus the precision matrix with fewer elements, making it quicker to compute this inversion and implement into a Neural Network. For a kernel $k(\boldsymbol{x}, \boldsymbol{y}) = k(\Delta)$ where $\Delta = \boldsymbol{x} - \boldsymbol{y}$, we can use Bochner's Theorem which states: "A continuous kernel $k(\boldsymbol{x}, \boldsymbol{y}) = k(\Delta)$ on $\mathbb{R}^D$ is positive definite if and only if $k(\Delta)$ is the Fourier transform of a non-negative measure." [Rudin 1990]. Using $p(\omega)$ as our general

non-negative measure, the Fourier transform is thus

$$k(\Delta) = \int p(\omega) \exp(i\omega\Delta) d\omega \tag{6}$$

where we can select the distribution $p(\omega)$.

Now consider the function $h(\boldsymbol{x}) = e^{i\boldsymbol{\omega}\boldsymbol{x}}$ where $\boldsymbol{\omega} \sim \mathcal{N}(0, 1)$ is a vector of normally distributed random values with probability density function $p(\omega)$. Taking the expectation value of $h(\boldsymbol{x})h(\boldsymbol{y})^*$ with respect to ω yields:

$$\begin{aligned} \mathbb{E}[h(\boldsymbol{x})h(\boldsymbol{y})^*] = & \int p(\omega) \exp\left(i\boldsymbol{\omega}^\top(\boldsymbol{x}-\boldsymbol{y})\right) d\boldsymbol{\omega} \\ = & \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{y})^\top(\boldsymbol{x}-\boldsymbol{y})\right) \end{aligned} \tag{7}$$

which is equivalent to an RBF kernel with $\sigma = 1$. Defining a random map $\boldsymbol{h}(x)$ as a normalised vector of R functions $h_r(\boldsymbol{x}) = e^{i\boldsymbol{\omega}_r\boldsymbol{x}}$ for $r \in [1, R]$ we now get

$$\boldsymbol{h}(x) = \frac{1}{\sqrt{R}} \begin{bmatrix} e^{i\boldsymbol{\omega}_1\boldsymbol{x}} \\ \vdots \\ e^{i\boldsymbol{\omega}_R\boldsymbol{x}} \end{bmatrix} \tag{8}$$

such that

$$\boldsymbol{h}(x)\boldsymbol{h}(y)^* = \frac{1}{R}\sum_{r=1}^{R} \exp\left(i\boldsymbol{\omega}_r^\top(\boldsymbol{x}-\boldsymbol{y})\right) \tag{9}$$

$$\approx \mathbb{E}[\exp\left(i\boldsymbol{\omega}_r^\top(\boldsymbol{x}-\boldsymbol{y})\right)] \tag{10}$$

$$= k(\boldsymbol{x}-\boldsymbol{y}) \tag{11}$$

where we have effectively taken R Monte-Carlo samples to approximate the expectation value of $h(\boldsymbol{x})h(\boldsymbol{y})^*$ using multiple random features ω_r. The product is complex however, and thus expensive to compute. We can make a further simplification to allow easier implementation and computation.

Since both the kernel and Normal distribution from which we draw ω are real-valued, we can consider only the real part of the complex exponential

$$e^{i\omega_r^\top(\boldsymbol{x}-\boldsymbol{y})} = \cos(\omega_r^\top(\boldsymbol{x}-\boldsymbol{y})) \tag{12}$$

using Eulers formula. Now defining $z_\omega(\boldsymbol{x}) = \sqrt{2}\cos(\omega^\top\boldsymbol{x} + b)$ with b randomly sampled from the interval $[0, 2\pi]$, we can write the expectation of the product $z_\omega(\boldsymbol{x})z_\omega(\boldsymbol{y})$ as:

$$\mathbb{E}[z_\omega(\boldsymbol{x})z_\omega(\boldsymbol{y})] = \mathbb{E}[\cos(\omega^\top(\boldsymbol{x}+\boldsymbol{y}) + 2b)] + \mathbb{E}[\cos(\omega^\top(\boldsymbol{x}-\boldsymbol{y}))] \tag{13}$$

where we have used the trigonometric identity for the product of 2 cosines. The first term on the RHS of Equation 13 will average to 0 since each phase term b is drawn from a random uniform distribution.

In a similar fashion to $\boldsymbol{h}(\boldsymbol{x})$, we can define a normalised vector of $z_{\omega_r}(\boldsymbol{x})$ values as:

$$\Phi(\boldsymbol{x}) = \sqrt{\frac{2}{R}} \begin{bmatrix} \cos(\omega_1^\top \boldsymbol{x} + b_1) \\ \vdots \\ \cos(\omega_R^\top \boldsymbol{x} + b_R) \end{bmatrix} \tag{14}$$

where each element is known as a random feature.

The covariance kernel is now approximated as $\Phi^\top(\boldsymbol{x})\Phi(\boldsymbol{x})$ which is an $N \times N$ matrix where N is the number of samples in a batch. Computing $\Phi(\boldsymbol{x})$ is cheap and allows us to calculate the GP mean and covariance via inversion of the precision matrix in Equation 5.

The two layers used in place of the final layer are:

$$logits(x) = \Phi(x)\beta, \qquad \Phi(x) = \sqrt{\frac{2}{R}} \cos(Wx + b), \tag{15}$$

where $\Phi(x)$ are the R Random Fourier Features, W is a set of R fixed weights drawn from a Gaussian $\sim \mathcal{N}(0, 0.05^2)$ and $b \in [0, 2\pi]$. β is a set of learnable parameters applied over the Random Features, and is essentially a fully connected layer with no bias term. As R increases, the approximation of the covariance matrix becomes more accurate as we take more samples and map out the Fourier space.

The logits and covariance matrix now define a Gaussian Distribution that can be sampled from. Traditionally this is done using Monte-Carlo samples, however a faster approach is the mean-field method [Lu, Ie, and Sha 2021]

$$\text{output}(x) = \text{softmax}\left(\frac{logits(x)}{\sqrt{1 + \lambda\sigma^2(x)}}\right) \tag{16}$$

where the variance $\sigma^2(x)$ is the diagonal element of the covariance matrix associated with the image x, and λ is known as the mean-field scalar that can be tuned in a similar fashion to the temperature.

2.5. Batch Normalisation

Batch Normalisation (BN) [Ioffe and Szegedy 2015] is a method to improve the sensitivity and convergence of neural networks by re-scaling the outputs of a given layer to have 0 mean and unit variance. For a layer with output dimension d and activations $\boldsymbol{x} = (x^{(1)}, x^{(2)}, ..., x^{(d)})$, we normalise each dimension independently,

$$\hat{x}_i^{(k)} = \frac{x_i^{(k)} - \mu_B}{\sqrt{\sigma_B^2}} \tag{17}$$

where μ_B and σ_B^2 are respectively the mean and variance of the activations for a given batch B and k is the dimension. The result of this is also further constraint to the Lipschitz constant of the network and is an alternative method of increasing network smoothness compared to spectral normalisation [Khan, Hayat, and Porikli 2017].

2.6. Uncertainties

We utilise 2 uncertainty metrics from our previous work [Hudson and Millicheap 2023], the predictive entropy $\mathbb{H}$ and the overlap index $\langle\eta\rangle$. The predictive entropy [Gal 2016] for an unseen

data sample x^* is defined as

$$\mathbb{H}(y|x^*, \mathcal{D}) = -\frac{1}{\ln(C)} \sum_{c}^{C} \left(q(y = c|x^*, \mathcal{D})\right) \ln\left(q(y = c|x^*, \mathcal{D})\right). \tag{18}$$

where C is the total number of classes. This can be extended to MC-dropout models by using the average softmax probability for each class. Additionally, for MC-dropout models we turn off weights at random in the final layer with probability $p_{MC} = 0.5$ and perform $N = 100$ forward passes through the model, thus providing a distribution of predictions with varying confidence levels. The overlap index $\langle\eta\rangle$ [Pastore and Calcagnì 2019] is computed from this distribution of softmax probabilities by assuming a Gaussian distribution for each class and then computing the overlapping area between them and is derived in Hudson and Millicheap 2023. It is unsuitable for deterministic models such as Gaussian Processes as these produce only 1 softmax output per image.

For Gaussian Processes we can take the diagonal elements of the covariance matrix to be the variance of each sample with respect to the training data. These values depend heavily on the kernel scale factor and are not bounded, meaning they are a poor choice as a measure of uncertainty for representing the error in a prediction. We therefore use the predictive entropy as the measure of uncertainty for the GP models, with the variance being accounted for in the mean-field approximation.

2.7. Calibration Metrics

In order to be perfectly calibrated, a model should output a probability of classification that directly equates to the rate of correct classification. Likewise, the magnitude of the uncertainty should correspond to the chance of incorrect classification. We follow on from our previous work [Hudson and Millicheap 2023] and use the Expected Calibration Error (ECE) and Uncertainty Calibration Error (UCE) metrics to quantify the degree of mis-calibration in our models,

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|acc(B_m) - conf(B_m)\right|, \tag{19}$$

$$UCE = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|uncert(B_m) - err(B_m)\right| \tag{20}$$

where we bin the n test images by confidence and uncertainty into M histogram bins of equal width, each denoted by B_m. $acc(B_m)$ is defined as the average rate of correct positive classifications, while $conf(B_m)$ is the average class prediction (0 to 1) given by the softmax probabilities. We compute the mis-classification rate, err, and use this to define $err(B_m)$ as $2\times\min(err, 1-err)$ such that an error of 1 equates to 50% model accuracy and thus should correspond to the maximum uncertainty. $uncert(B_m)$ is the mean uncertainty measure (average overlap index, predictive entropy) in the bin. The ECE and UCE are commonly expressed as percentages, with 0% corresponding to perfect calibration.

3. Models

In this work we utilise the models from our previous work [Hudson and Millicheap 2023], which consist of 1 non-equivariant LeNet-5 based model [Lecun et al. 1998], 3 cyclically equivariant models

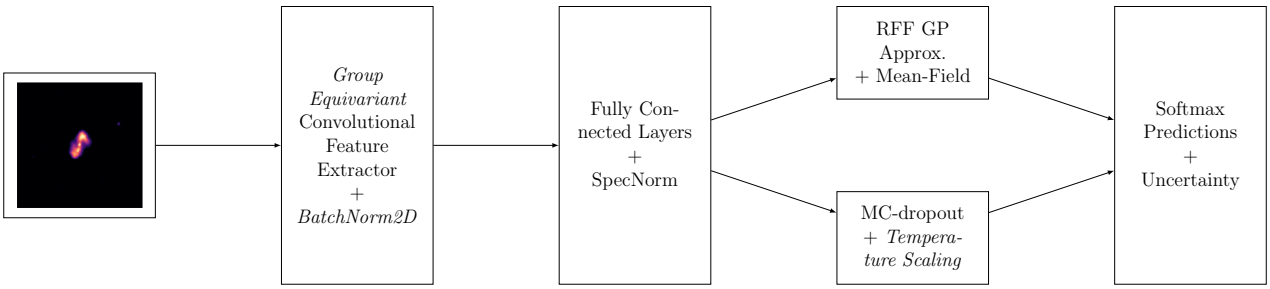


Figure 1: Flow diagram demonstrating the model architecture with 2 possible classification methods. Operations in italics are optional and are applied in different scenarios.

of 4, 8 and 16-fold rotational symmetry and 3 dihedrally equivariant (rotations and reflections) models of the same orders. We focus only on modifying the final layer of the models. The final 84-to-2 neuron linear layer was replaced with a 1024 RFF layer with fixed random weights and biases drawn from the distributions in section 2.2 and 1024 learnable parameters. We also introduced Batch Normalisation to each convolutional layer and Spectral Normalisation after each linear layer to ensure smoothness. The mean-field approximation is then applied to the model output in order to approximate a multivariate Gaussian distribution. A brief diagram of the architecture is displayed in Figure 1. In order to simplify comparisons between models we chose to modify and work with a smaller subset of just 3 models, the non-equivariant model, the lowest order equivariant model C_4 and the highest order of equivariance D_{16}. It is suggested that for a Gaussian Process attached to a Deep Learning model to perform well, the preceding network must be both smooth and sensitive to any inputs [Amersfoort et al. 2022]. Smoothness is achieved through the use of SN on all fully-connected layers, while sensitivity means that for a small but significant change in the input, the model is sufficiently able to detect this and is implemented with BN.

3.1. Training Process

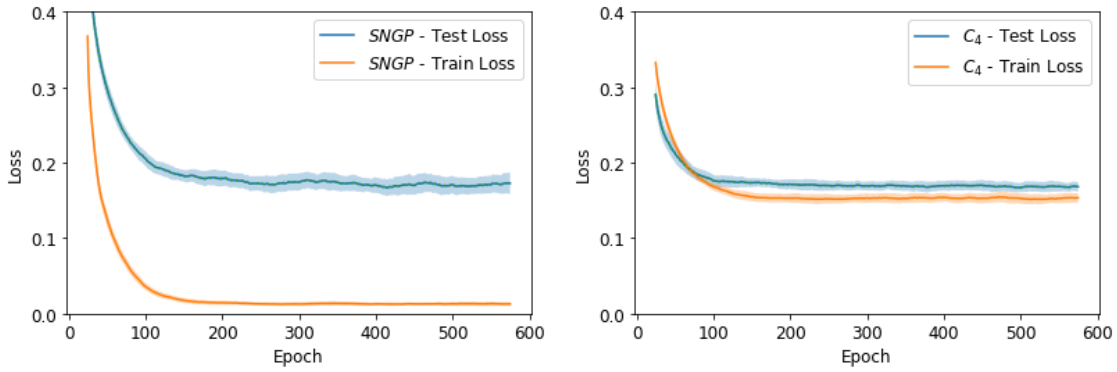


Figure 2: Training and validation loss during training of the SNGP model compared to the C_4 model averaged over 3 training runs and smoothed over 20 epochs. Both models achieve a similar validation loss, however the SNGP shows significant divergence between test and train loss.

Models were trained on the MiraBest Confident dataset [Miraghaei and P. N. Best 2017] in a similar fashion to our previous work. All models were trained for 600 epochs using an Adam optimiser with a learning rate of 10^{-4} for MC-dropout models and 10^{-3} for GP models, with a mini-batch size of 128. Early stopping based on the validation accuracy was used to prevent over-fitting. During training the GP models we found it necessary to implement 2D Batch Normalisation

to each convolutional layer to achieve competitive accuracy with the MC-dropout models. This resulted a large increase in validation accuracy from around 90% and also allowed for the learning rate to be increased from 10^{-4} to 10^{-3}, resulting in quicker convergence to similar performance as the dropout-enabled models.

After training the GP models, it was noticed that the training and test loss quickly diverged as displayed in Figure 2. This is in direct contrast to the loss curves for our original models in which train and test loss remain close in value. At first glance this may suggest extreme over-fitting to the training data, however the regularisation methods applied should help to mitigate this. Since we use the cross-entropy loss computed between the predicted and assigned labels, the number of misclassifications during training has a large impact on the loss function and explains the discrepancy observed. The GP models consistently scored 98-99% accuracy on the training data with a loss value averaging 0.03 whereas MC-dropout models remained in the range of 95% accuracy with a loss value in the range of 0.15. When applied to a reserved test set, the model scored better than all previous models at 98.1%, implying that there was little to no loss of generalisability due to over-fitting.

4. Comparative Analysis of Calibration Methods

Table 1 compares the non-equivariant LeNet model to the lowest (C_4) and highest (D_{16}) orders of equivariance with additional modifications displayed below the main model. The following subsections discuss each modification separately and its effects on the performance.

Model	**Accuracy**	**ECE**	**UCE** $\langle\eta\rangle$	**UCE** $\mathbb{H}$
LeNet+SN	94.5%	5.58%	1.30%	33.34%
+BN	91.1%	3.89%	0.55%	26.77%
+TS	94.5%	**0.33**%	0.42%	10.25%
+BN+GP	98.1%	1.79%	//	**5.43**%
C_4+SN	94.9%	4.10%	0.83%	27.57%
+BN	94.4%	3.44%	0.98%	23.88%
+TS	94.9%	0.58%	0.38%	11.32%
+BN+GP	**99.1**%	1.44%	//	7.51%
D_{16}+SN	95.7%	3.78%	0.70%	25.39%
+BN	95.1%	2.89%	0.77%	22.30%
+TS	95.7%	0.44%	**0.25**%	7.72%
+BN+GP	95.2%	2.99%	//	7.64%

Table 1: Ablative comparison of model modifications and their impact on accuracy and calibration scores. All models are spectrally normalised by default, we then apply batch normalisation (BN), temperature scaling (TS) and finally replacing the MC-dropout classifier with an RFF GP approximation with BN applied. $\langle\eta\rangle$ denotes the overlap index as the uncertainty measure (not applicable for deterministic models), while $\mathbb{H}$ is predictive entropy.

4.1. Batch Normalisation

Although we found it necessary to include batch normalisation when using the GP classifier head, implementing this technique to the MC-dropout enabled models resulted in worsened predictive performance in all cases while increasing calibration only slightly. This is most notable for the LeNet model which experienced a 3% loss in predictive accuracy compared to its unmodified counterpart. Li et al. 2018 consider this and suggest that applying both MC-dropout and

BN together can cause variance shifting in the trained models. Since BN controls the variance of neuron activations causing them to remain as close to unity as possible, dropping-out 50% of neurons in any layer can cause an unprecedented shift in this variance, leading to poorer predictive performance.

It is a matter of debate whether to apply batch normalisation prior to or after the activation function. Both approaches were tested and it was found that applying batch normalisation before the ReLU function resulted in a less expressive latent space representation using the GP LeNet model, and validation accuracy rarely rose above 90%. After swapping the order of operations, we were able to achieve an accuracy of 98.1%. We suggest that this is due to the pruning effect of the ReLU function, if it were applied after normalisation to mean 0 and variance 1, approximately half of all latent values will be negative and thus set to 0 by the ReLU function, reducing the expressivity and losing a large amount of information in the process. Contrary to Ioffe and Szegedy 2015, it has also been found empirically [Chen et al. 2019] that applying BN after the activation function leads to improved results.

4.2. Temperature Scaling

By considering the un-scaled logits prior to the application of the softmax function, we were able to minimise the ECE over a range of temperature values for each model. We found that all orders of equivariance required a similar optimum temperature of ≈ 2.5 to achieve good calibration. Figure 3 shows the ECE plots for 3 models before and after scaling while Figure 4 displays the UCE plots for the same 3 models. It can be seen that the sigmoid shape of the ECE plots is no longer present after temperature scaling, indicating almost perfect calibration and the removal of any under-confidence.

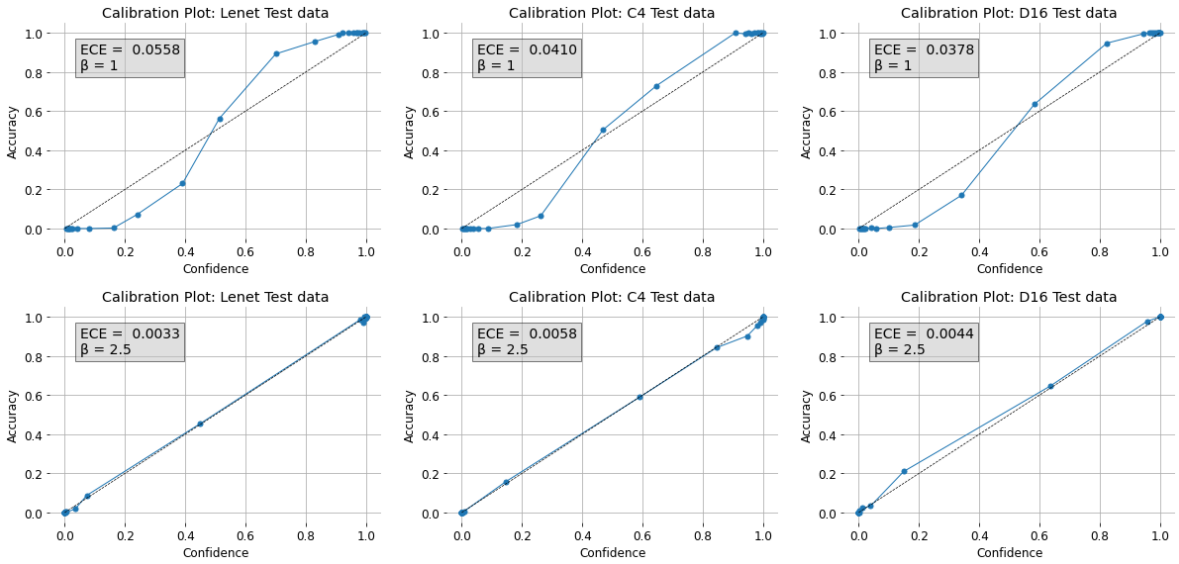


Figure 3: Calibration plots for the LeNet model before and after temperature scaling was applied with $\beta = 2.5$. All plots use M=26 histogram bins. Scaled ECE values were all lower by an order of magnitude. The sigmoid shape was flattened, showing the models are no longer under-confident.

Figure 4 displays UCE $\langle \eta \rangle$ plots using the average overlap as a measure of uncertainty for 3 models before and after temperature scaling. Since the overlap index was already a well-calibrated measure of uncertainty [Hudson and Millicheap 2023], the reduction in the UCE was not as extreme as the ECE measure, however all models experienced a decrease with the D_{16} model having the

lowest UCE of just 0.25%. Also notable is that the mean assigned uncertainty value was reduced by temperature scaling which can be seen in the slight shifting of bins to the left in Figure 4. The ECE is greatly affected by the number of bins used to plot it, however once calibrated we discovered that the variation due to the number of bins was negligible.

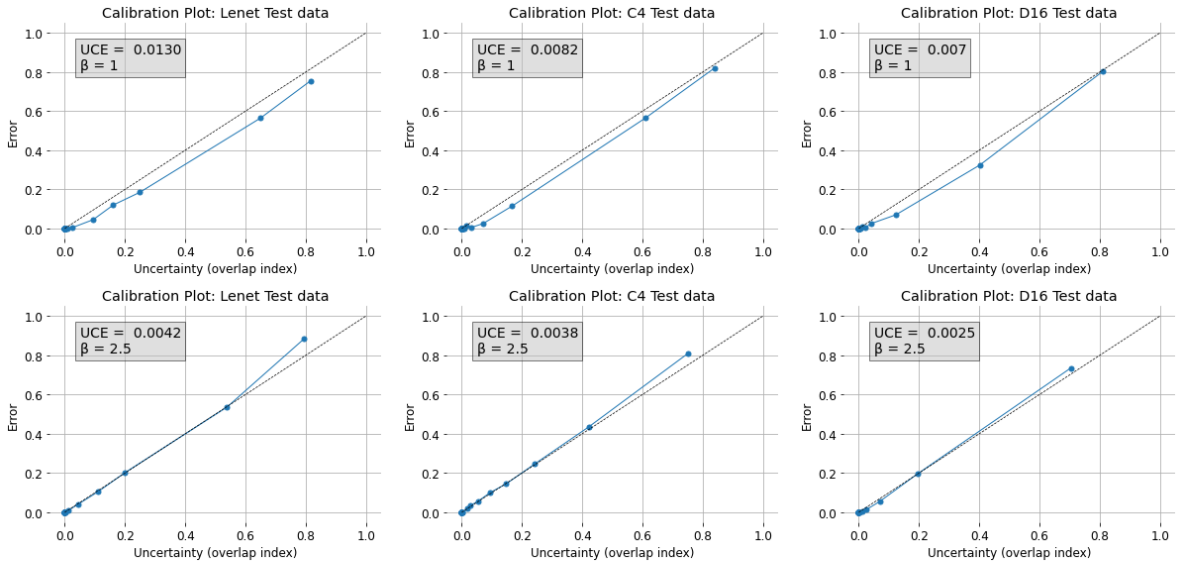


Figure 4: Calibration plots for 3 models before and after temperature scaling was applied with $\beta = 2.5$. All plots use M=26 histogram bins. Scaled UCE values were all reduced, while mean uncertainty was also reduced, suggesting less assigned uncertainty.

The predictive entropy as a measure of uncertainty for MC-dropout models is still poorly calibrated with a minimum UCE $\mathbb{H}$ achieved for the D_{16} model of 7.72%, significantly larger than the minimum of 0.25% using the overlap index. Since the entropy is a non-linear logarithmic function of the softmax probabilities and is compared against the error (a linear function of accuracy), it becomes impossible to have a perfectly calibrated ECE and UCE $\mathbb{H}$ at the same time for any probability $\neq 1$. It would therefore be preferable for future work to investigate an alternative measure of uncertainty for GP models that can be calibrated independently of the softmax probabilities.

4.3. GP Calibration

Attaching the GP head to the standard LeNet model using a non-equivariant feature extractor was able to achieve 98.1% accuracy on the reserved test set, while our best model previously was found to be the D_4 model with an accuracy of 97.5% on the same dataset [Hudson and Millicheap 2023]. This trend continues with the C_4 model achieving 99.1% accuracy on average which is our best performing model to date. We were unable to see a similar improvement with the larger D_{16} model as it had to be trained on a server, time constraints meant that it was not able to be subjected to the same amount of hyper-parameter tuning as our other models, and as such the results for this model are not considered further.

The ECE of the GP models is better than the uncalibrated MC-dropout models, but falls short of the temperature scaled models. One possible reason for apparent improved calibration is the greater predictive accuracy of these models such that more points lie at the extreme ends of the ECE plots, which has a minimising effect. On the other hand, since the GP models provide a single deterministic output, the calibration metrics suffer from discretisation of the bins leading to poorer calibration scores. To properly probe the calibration of the GP models further, we would

require a large dataset of more uncertain sources or an ensemble of models. From Table 1 the best UCE $\mathbb{H}$ was achieved with the GP LeNet model at 5.43%. This is still significantly larger than the UCE $\langle\eta\rangle$ of the calibrated models, indicating that the uncertainties from the GP model could benefit from further tuning and calibrating.

It was noticed that greater variance values of most samples did not necessarily line up with larger assigned entropy by the model, indicating that the variance better represents aleatoric rather than epistemic uncertainty. We also compared the variance between the LeNet and C_4 models, finding that a similar subset of images were assigned the highest variances by both models, again suggesting this measure is aleatoric in origin rather than due to the training of the models.

There is empirical evidence that increasing the size of a neural network can improve performance without over-fitting to the training data if there is sufficient regularisation (Batch and Spectral Normalisation in our case) applied [Krizhevsky, Sutskever, and Hinton 2012]. We suggest that the increased performance of the GP models could in part be due to the additional 1024 learnable parameters introduced into the model, raising the learning capacity. Experiments with fewer random features while keeping other hyper-parameters fixed resulted in inferior performance. The C_4 model with 256 features scored 96.2% on the reserved test set compared to 99.1% with 1024 features. It is important to note that the number of features also greatly impacts the accuracy of the covariance matrix approximation which can then have an impact on the accuracy such that it is difficult to disentangle these two factors.

5. Implications for Astrophysical Analysis

5.1. Detecting Out of Distribution Data

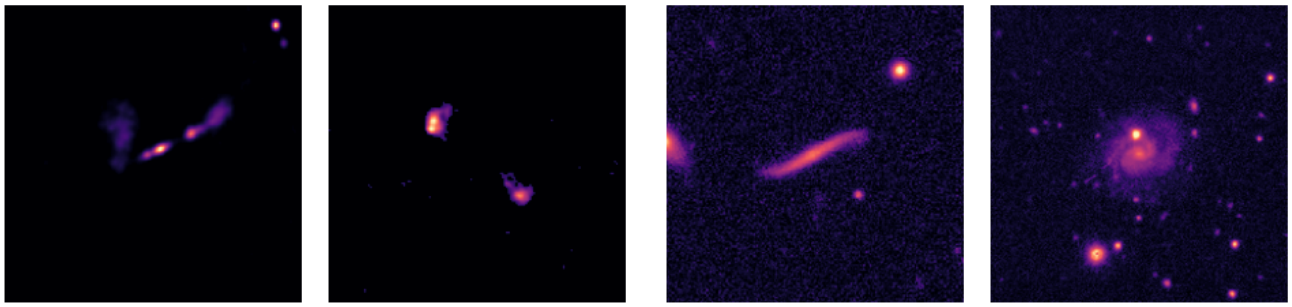
An important test of neural networks is the ability to accurately determine whether an unseen image is actually part of the classification problem or completely unrelated. In our previous report [Hudson and Millicheap 2023] we discussed the possibility for Spectral Normalisation to improve Out of Distribution (OoD) detection capabilities [Miyato et al. 2018]. In this section we test our models with different classifiers to explore how these images are processed within the model and the effects of different feature extractors and classifiers.

We selected additional astronomical images to use as OoD data, in particular the optical Galaxy MNIST[1] dataset, which is comprised of 10,000 images of galaxies as seen in the optical spectrum with 4 labelled classes. The images are presented in 2 formats, a 64×64 and a higher-resolution of 224×224. Following a similar method of preparing the images to that of Aniyan and Thorat 2017, we first cropped the 224×224 images to 150×150 centered on the middle. Since 3 RGB channels were provided, it was decided that only the red channel was to be used to be consistent between images. The images were then normalised to have pixel values between $0 - 1$ using the formula

$$\text{New Image} = \frac{\text{Image} - min(\text{Image})}{max(\text{Image}) - min(\text{Image})}. \tag{21}$$

To combine these images into our test data-set of 104 FR type radio galaxies, we selected a random sample of 50 images from GalaxyMNIST such that we would have an approximately equal number of Optical, FRI and FRII images. Figure 5 presents some optical galaxies next to various examples of FR-type radio galaxies. Pre-processing the GalaxyMNIST images excluded any form of noise reduction techniques, resulting in fairly noisy backgrounds. Further work could apply proper noise reduction techniques to these images.

[1] https://github.com/mwalmsley/galaxy_mnist



(a) FR I (left) and FR II (right) radio galaxies

(b) Optical GalaxyMNIST galaxies

Figure 5: 2 examples of galaxies from the MiraBest Confident test set against 2 out of distribution optical galaxy images from the Galaxy MNIST dataset.

We can extract the values in the latent space between any layers in the neural network, and can use these to gather information on what is happening within the model without being restricted to just the input and output. The latent space prior to the softmax function can be directly plotted along with the models assigned uncertainty to visualise the decision boundary between classifications and provide a possible way to identify regions that OoD data is mapped to. As shown in Figure 6 the OoD data was centered about (0,0) while extending into a much larger proportion of the latent space for the GP classifier, whereas the MC-dropout model was able to constrain this data much more tightly around the origin. The GP model has greater separation between the two ID classes and there are correspondingly very few galaxies in the same region as the OoD data, it is therefore difficult to determine which model is better at separating this from the ID data. Although the OoD data was indeed being identified by the models, the GP is seemingly unable to express this properly in its final outputs by assigning high uncertainty to it, contrary to what we expected. The kernel scale hyper-parameter controls the 'distance' between a new image and the training data, reducing this scale could make the drop-off from ID to OoD data sharper and thus more easily identifiable by concentrating the data about the origin.

To isolate this data, a region in the latent space of the model could be selected and any images mapped to this region flagged for further inspection. For example in the case of the MC-dropout model in Figure 6, any point with a distance magnitude of less than 2 from the origin would successfully isolate all of the OoD images while only flagging a few of the most uncertain sources. Further work could push the limits of OoD detection by using much larger catalogues of varying images to map out more of the latent space.

To explore the models OoD detection capabilities further, we can consider the high-dimensional latent spaces before any classification is performed. One way to express this high-dimensional data in fewer dimensions is through a Uniform Manifold Approximation and Projection (UMAP) [McInnes, Healy, and Melville 2020]. UMAP preserves the connections between datapoints in the high-dimensional data and this is then reflected in the distances between points in the lower-dimension expression. The projection utilises Gaussian random walks and thus is not reproducible, however one example projection is shown in Figure 6 where we were able to observe 3 distinct regions in the mapping.

While the absolute distance is not important, the relative separation of each class in UMAP space is meaningful, allowing us to draw conclusions on the OoD detection capabilities of different feature extractors (Equivariant orders). The ratio of the distance to the OoD class from the centroid of the ID classes to the distance between the ID classes was computed over 5 different UMAPs to be 1.5 ± 0.2 for the non-equivariant model and 3.07 ± 0.7 for the C_4 feature extractor.

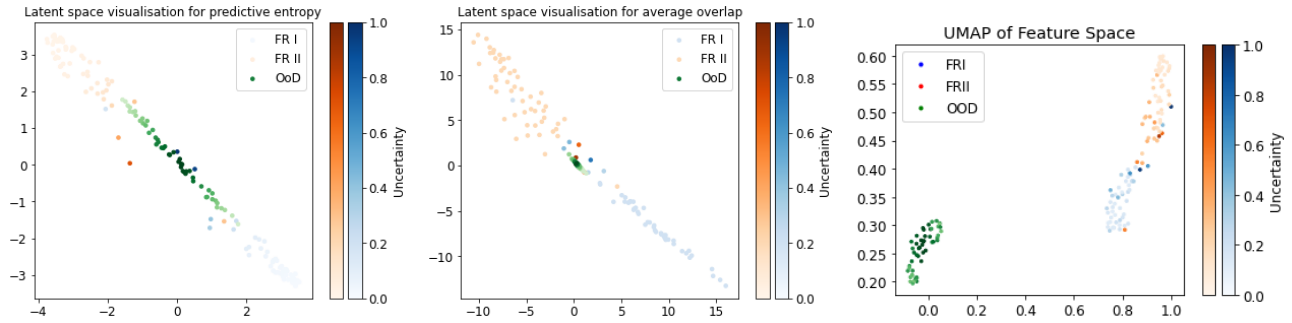


Figure 6: UMAP of the 84-dimensional space prior to the classifier using the LeNet feature extractor (left) with the latent spaces of the output layer prior to the softmax function for different classification techniques (centre and right). (centre) GP classifier with uncertainties represented by predictive entropy. (right) MC-dropout classifier with temperature scaling applied, uncertainty is measured in average overlap.

This suggests with high confidence that higher-order equivariance models naturally lead to better OoD data detection since it is on average 2 times further from the ID domain and thus will be easier for the classifier to distinguish it.

5.2. Luminosity-Size Distinction

Understanding the origin of the model assigned uncertainties is crucial to being able to use them in any meaningful capacity. Historically, the distinction between FRI and FRII type galaxies could be made using the luminosity boundary of 10^{25} WHz^{-1}, although surveys with more sensitive instruments have revealed a large population of fainter sources [Mingo et al. 2019].

We utilised the integrated luminosity and angular size data from the original MiraBest paper [Miraghaei and P. N. Best 2017] to create a plot of luminosity against actual size in kpc for the galaxies in the MiraBest Confident dataset. To convert from arseconds to kpc we note that all galaxies in the original MiraBest set have observed redshifts of $z < 1.5$. Thus we use the Hubble Law ($H_0 = 70$ kms^{-1}Mpc^{-1}) approximation to calculate the actual size since the distance-redshift relation can be modelled as a linear function at low redshifts. We also use a similar redshift cut-off of $0.01 < z < 0.8$ as Mingo et al. 2019, however found that this excluded no sources from the test dataset.

Maximally separating the two classes is done using a support vector classifier (SVC), this yields a a boundary line with uncertainties that we can use to compare between various classifying techniques. The baseline boundary is found by fitting a SVC to the full MiraBest Confident dataset using the human-assigned labels and no uncertainty. The equation of this boundary line is $m = 1.40 \pm 0.17$ and $c = 22.06 \pm 0.35$. To ensure that the reserved test set is representative of the full MiraBest population we compared the dividing line and found that they were in near complete agreement, with a test set boundary of $m = 1.42 \pm 0.6$ and $c = 22.06 \pm 1.2$ confirming that the test set is a good representation of the full MiraBest population.

Immediately apparent in the MiraBest dataset is the tendency for larger, fainter sources to be FRI, while smaller but more luminous sources are FRII as shown by the relative positions of the medians in Figure 8. This is in direct contrast to the findings of Mingo et al. 2019 in which the LoTSS dataset FRII sources were larger and brighter on average. We also note that there is a distinct lack of high-luminosity FRI sources above 10^{26}WHz^{-1} in our data and a significant portion of FRII sources lie below the canonical FR break of 10^{25}WHz^{-1} resulting in a significant overlap in the populations. These differences can be partly explained by using different telescopes at different frequencies. MiraBest sources were imaged using the VLA telescope at a frequency of 1.4

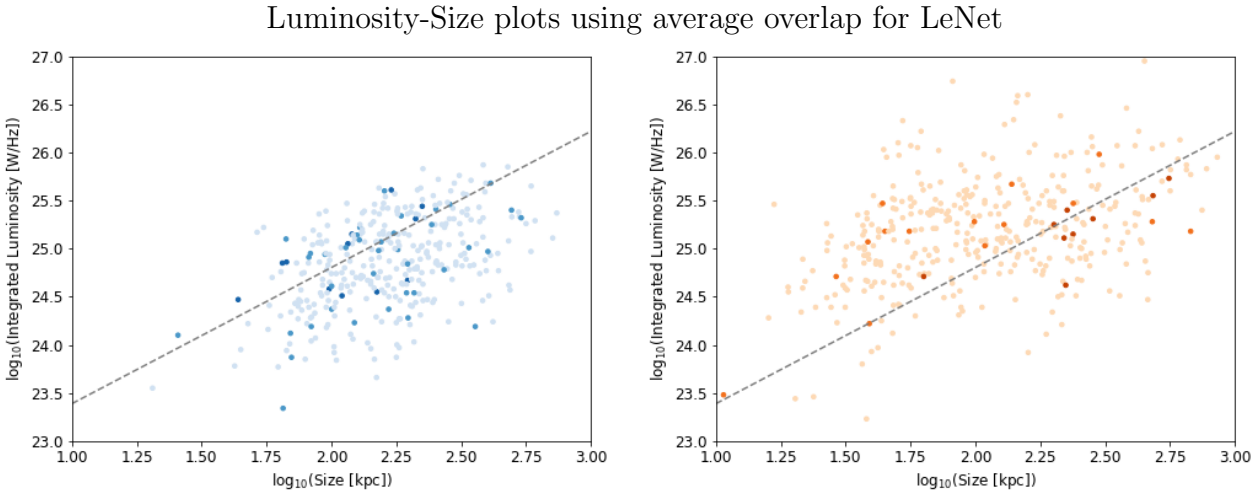


Figure 7: Logarithmic plot of integrated luminosity against actual size of all radio galaxies in the MiraBest Confident training set with a maximally dividing separation line of $m = 1.40 \pm 0.17$ and $c = 22.06 \pm 0.35$. Split by class for clarity with FRI galaxies in blue, FRII in orange, with more intense points representing higher model assigned uncertainty. Uncertainties are not used in calculating the fit and are shown to demonstrate the regions where high uncertainty is assigned

GHz whereas the LoTSS dataset utilises LOFAR at 150 MHz. Morganti, Killeen, and Tadhunter 1993 found that the overlapping region of FR galaxies is much wider at higher frequencies. The VLA is also capable of higher resolutions than LOFAR, thus explaining why the sizes of the MiraBest galaxies are roughly an order of magnitude less than those in LoTSS. Differences in the two populations may also be a result of the different selection techniques that were applied to the datasets, for example Miraghaei and P. N. Best 2017 exclude various sources such as ones larger than the 150×150 image size.

For each model tested we instead use the model predicted classes rather than the true labels while each galaxy is weighted by $1/\sigma^2$ where σ is the assigned model uncertainty. To prevent extremely confident sources from dominating and skewing the resulting fit, we apply a cut-off such that all uncertainties < 0.2 are treated as confidently classified and set to 0.2. Figure 7 displays the distribution of each class of MiraBest galaxies on the luminosity-size axes with the baseline dividing line. Uncertainties from the LeNet model are included to indicate that higher uncertainties are loosely clustered in the overlapping area of the two classes. We then present Figure 8, showing the smaller test set distribution of luminosity-size for different models using their best-calibrated uncertainties. Alongside each plot is a histogram of the distribution with the median marked.

Comparing the gradient and intercept of the class division using different models can provide insights into the nature of the classifiers. Using the assigned predictions from GP models, we found the best boundary to be $m = 1.45 \pm 0.6$, averaged over the LeNet, C_4 and D_{16} models. We can attribute the similarity in the FR break to the baseline to the strong predictive power of these models, with only 2, 1 and 5 mis-classifications respectively. MC-dropout models provided a steeper gradient, with an average of $m = 1.86 \pm 0.6$. With such a small data-set of only 104 images it is difficult to make any concrete conclusions about these results due to the high uncertainties and sparsity of the data, however the steeper gradient of the MC-dropout classifiers suggests that the machine learning models are more likely to assign FRI to larger, brighter sources.

Looking at Figure 8 the number of FRI sources incorrectly labelled as FRII by the models (red crosses, 11 instances) is far greater than that of mis-classified FRIIs (blue crosses, 5 instances). Using other models we also observed a similar trend of more mis-classified FRI galaxies and the

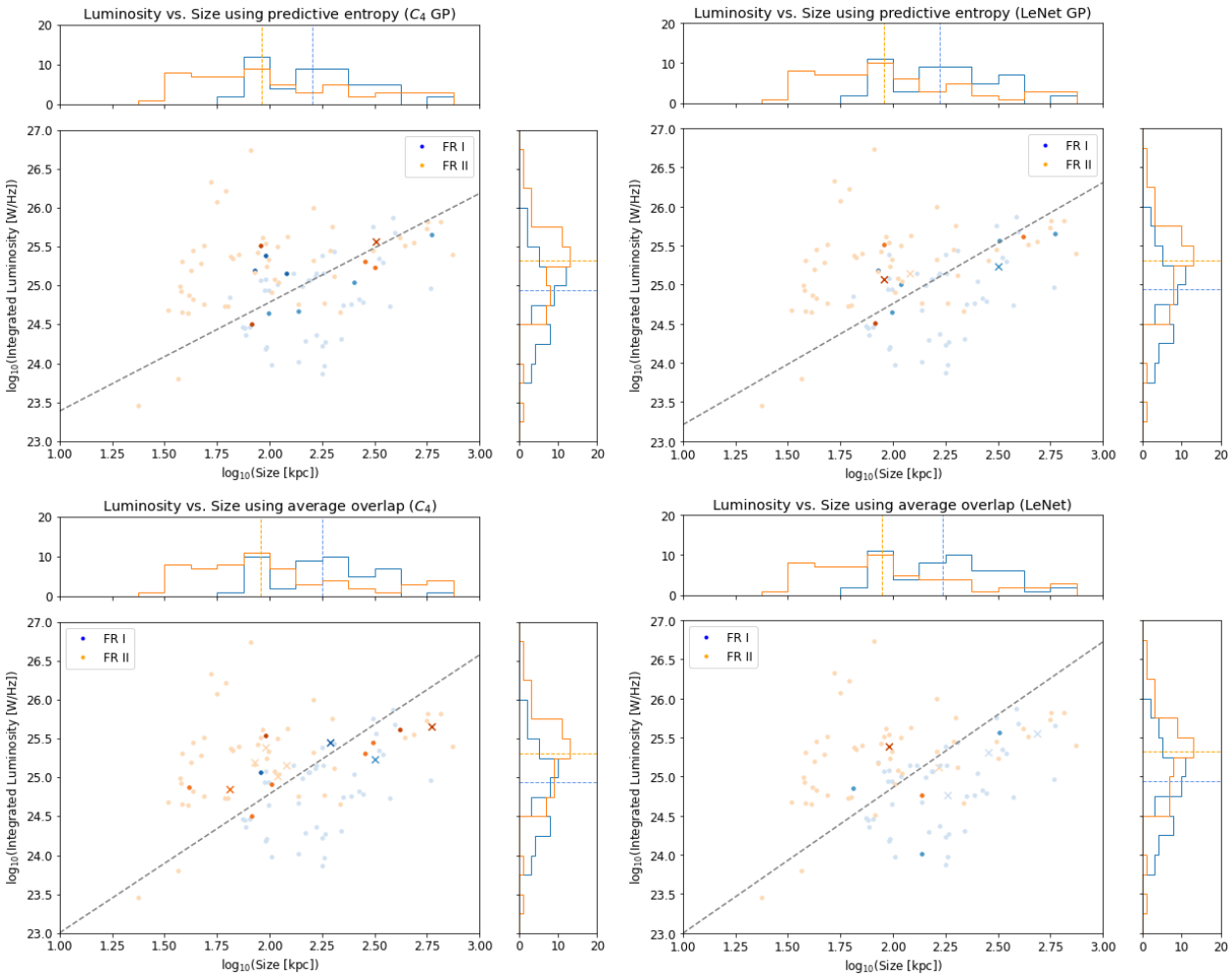


Figure 8: Comparison of C_4 (left) and LeNet (right) models with different classification methods applied to the reserved test set. $\times$ markers highlight any mis-classified sources such that a blue $\times$ denotes an FRII source mis-identified as FRI by the model, while colour intensity represents uncertainty. (top) GP classifiers with assigned classes and uncertainties shows a best fit of $m = 1.40 \pm 0.6$ and $c = 22.0 \pm 1.3$ (C_4) and $m = 1.50 \pm 0.7$ and $c = 21.66 \pm 1.5$ (LeNet). (bottom) Temperature scaled MC-dropout classifiers with a best fit of $m = 1.79 \pm 0.7$ and $c = 21.2 \pm 1.4$ (C_4) and $m = 1.86 \pm 0.5$ and $c = 21.1 \pm 1.1$ (LeNet). Histograms display the median size/luminosity for each class with a dashed line.

mean assigned uncertainty is consistently higher for FRI sources. There exist many subgroups of FRI galaxies such as wide-angle tail sources and double-double morphologies [Mahatma, V. H. et al. 2019] resulting in more variance within the class. On the other hand FRII galaxies may be better classified due to edge-brightening, making it easier for the model to distinguish the necessary features. Both Scaife and Porter 2021 and Mingo et al. 2019 found a similar trend with FRII predictive accuracy noticeably better than that of FRI.

A visual inspection of the plots also shows that the majority of mis-classified and highly uncertain sources lie close to the dividing line and in the overlapping region between the classes. Since the models do not know the true size of the galaxies, only their angular size, the fact that mis-classified and uncertain sources lie close to this boundary in this parameter space indicates that there the exists a link between the underlying physical properties of the radio galaxies in this region and their more ambiguous morphologies. The use of confident human assigned labels introduces epistemic uncertainty into the classifications and may also explain why there is a higher

degree of uncertainty in this region.

6. Limitations and Further Research

6.1. Model Fine-Tuning

Implementing the Gaussian Process introduced a variety of hyper-parameters that required extensive tuning, such as the kernel length scale, mean-field factor and the covariance ridge penalty which controls the weighting of each mini-batch on the overall covariance matrix. There exists methods of optimising some of these hyper-parameters, in particular the kernel scale and mean-field factor can be optimised during training by introducing additional terms into the loss function and would be an important implementation for future work. Instead, we tuned these parameters manually in our experiments which may have lead to less than ideal results.

Adlam, Snoek, and Smith 2020 found a relationship between the use of Gaussian Processes and the prevalence of the Cold Posterior Effect (CPE) in which the amplitude of the kernel must be tempered to make good predictions. They also discuss the relationship of the CPE to the estimation of the aleatoric uncertainty σ^2_{noise} in equation 5. In terms of our model this would mean we also need to tune σ^2_{noise} to best represent the aleatoric noise in the training data to avoid the CPE. This may explain why we were unable to see well calibrated probabilities and uncertainties with the GP models since we used a fixed value of 0.01.

6.2. Data-set and Galaxy Properties

While we only looked at the luminosity and size of these galaxies, there are many more physical properties that are of interest to compare using the MiraBest set. The optical luminosity is known to be related to the radio luminosity [Ledlow and Owen 1996] and also provides a link to the super-massive black hole mass of the AGN. This would allow us to further probe the relationship between the host properties and FR classification.

Although care was taken to remove human-designated uncertain sources from the dataset, the fact that multiple sources were consistently mis-classified and marked as uncertain by different models suggests that these morphologies are somewhat different from the ID classes. A visual inspection in Figure 9 shows that the leftmost image is undoubtedly an FRI. The two central images share resemblances, yet belong to different classes. The center left appears to be a bent-tail morphology rather than a standard FRI, and shows significant edge-brightening. While the center right image shows edge-brightening, it has a central bright spot that many of the other FRII galaxies do not have. The rightmost image is in fact the only labelled as a Double-Double FRII morphology [Mahatma, V. H. et al. 2019], of which there are few in the training data, explaining its high uncertainty.

The introduction of another class may allow for separate OoD classification along with identifying sources of interest such as those in Figure 9, however this would require additional training and thus carefully chosen examples of sub-classes and labelled OoD data. Extending the problem to multiple classes may only confuse matters with many intra-class probabilities, so it would be preferable to instead assign these sources a sufficiently large uncertainty as they may be valuable to study on their own. Perhaps additional models trained on each broader class could provide more specific sub-classifications.

7. Conclusions

In this report we have showed that implementing a GP to the final layer of a smoothed convolutional neural network offers better speed and predictive performance over simpler MC-dropout

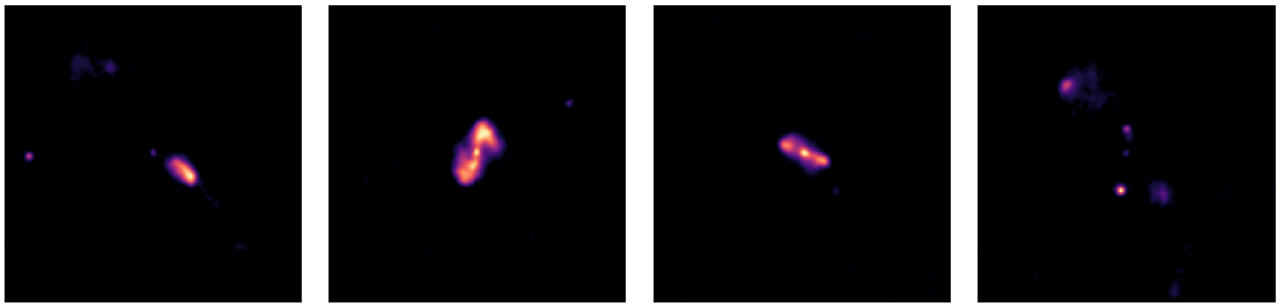


Figure 9: Consistently mis-classified and high uncertainty examples from the MiraBest Confident test set. (left) FRI sources and (right) FRII sources.

models, at the cost of less representative uncertainty measures. There is room for improvement, with further hyper-parameter tuning necessary to choose the ideal kernel length scale and mean-field factor. In addition, Batch Normalisation was successfully implemented into the convolutional layers of the GP models and resulted in a significant increase in accuracy while allowing the learning rate to be increased. It was then discovered that the interplay between BN and MC-dropout resulted in worsened performance due to variance shifting and the two techniques should not be used in tandem.

Although it was found that introducing a Gaussian Process improved the accuracy of the non-equivariant and C_4 classifier by a large margin, we were not able to conclusively determine which model was better at detecting OoD data. The MC-dropout models were able to isolate the OoD data to a smaller proportion of the full latent space while retaining expressiveness for the in distribution regions.

Applying temperature scaling to the logits as a post-training calibration method has been demonstrated to be a fast and efficient way of optimising the probabilities and uncertainties of relatively small deep learning models. We achieved almost perfect calibration for MC-dropout models with an expected error of just 0.25% using the overlap index, such that the model assigned uncertainties successfully quantify the chance of mis-classification. We failed to see good calibration using the predictive entropy and discussed how it is non-linearly linked to the softmax probabilities and is therefore a poor choice of uncertainty measure when trying to minimise both the ECE and UCE together.

Plotting the luminosity against size for the full MiraBest dataset showed that there was a large amount of mixing in the populations of radio galaxies. The canonical FR luminosity break found by Fanaroff and Riley was a poor fit for the MiraBest dataset, instead we required a significant portion of small, low-luminosity FRII galaxies and high-luminosity FRI sources to be accounted for. Our machine learning models consistently assigned high uncertainty and frequently mis-classified in this region of overlap and in close proximity to the boundary between classes, demonstrating a link between FR galaxies with similar physical properties having more ambiguous morphologies.

Overall, we have demonstrated the usefulness of uncertainties in a scientific application and explored various methods for calculating and calibrating them, with a focus on astrophysics. However the applications of machine learning are so widespread that these techniques are likely to have uses in other domains that utilise neural networks and require careful labelling of images.

References

Abazajian, Kevork N., Jennifer K. Adelman-McCarthy, and Agüeros (June 2009). "The Seventh Data Release of the Sloan Digital Sky Survey". In: 182.2, pp. 543–558.

Adlam, Ben, Jasper Snoek, and Samuel L. Smith (2020). *Cold Posteriors and Aleatoric Uncertainty*.
Amersfoort, Joost van et al. (2022). *On Feature Collapse and Deep Kernel Learning for Single Forward Pass Uncertainty*.
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.
Baum, Stefi A., Esther L. Zirbel, and Christopher P. O'Dea (Sept. 1995). "Toward Understanding the Fanaroff-Riley Dichotomy in Radio Source Morphology and Power". In: 451, p. 88.
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.
Chen, Guangyong et al. (2019). *Rethinking the Usage of Batch Normalization and Dropout in the Training of Deep Neural Networks*.
Cohen, Taco S. and Max Welling (2016). "Group Equivariant Convolutional Networks". In.
Condon, J. J. et al. (May 1998). "The NRAO VLA Sky Survey". In: *The Astrophysical Journal* 115.5, pp. 1693–1716.
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.
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*.
— (2016). *Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference*.
Gibbs, M.N. and D.J.C. Mackay (2000). "Variational Gaussian process classifiers". In: *IEEE Transactions on Neural Networks* 11.6, pp. 1458–1464.
Guo, Chuan et al. (2017). *On Calibration of Modern Neural Networks*.
Hudson, S. and A. Millicheap (2023). "Uncertainty calibration for group-equivariant Bayesian CNNs in radio galaxy classification". In: *MPhys Report*.
Ioffe, Sergey and Christian Szegedy (2015). "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In.
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. and Philip N. Best (2007). "Luminosity function, sizes and FR dichotomy of radio-loud AGN". In: *Monthly Notices of the Royal Astronomical Society* 381.4, pp. 1548–1560.
Khan, Salman, Munawar Hayat, and Fatih Porikli (2017). *Regularization of Deep Neural Networks with Spectral Dropout*.
Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc.
Laves, Max-Heinrich et al. (2019). *Well-calibrated Model Uncertainty with Temperature Scaling for Dropout Variational Inference*.
Lecun, Y. et al. (1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324.
Ledlow, Michael J. and Frazer N. Owen (July 1996). "20 CM VLA Survey of Abell Clusters of Galaxies. VI. Radio/Optical Luminosity Functions". In: 112, p. 9.
Li, Xiang et al. (2018). "Understanding the Disharmony between Dropout and Batch Normalization by Variance Shift". In.
Liu, Jeremiah Zhe et al. (2020). "Simple and Principled Uncertainty Estimation with Deterministic Deep Learning via Distance Awareness". In: *CoRR* abs/2006.10108.
Lu, Zhiyun, Eugene Ie, and Fei Sha (2021). *Mean-Field Approximation to Gaussian-Softmax Integral with Application to Uncertainty Estimation*.
Mahatma, V. H. et al. (2019). "LoTSS DR1: Double-double radio galaxies in the HETDEX field". In: *A&A* 622, A13.
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.
McInnes, Leland, John Healy, and James Melville (2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*.
Mingo, B et al. (July 2019). "Revisiting the Fanaroff–Riley dichotomy and radio-galaxy morphology with the LOFAR Two-Metre Sky Survey (LoTSS)". In: *Monthly Notices of the Royal Astronomical Society* 488.2, pp. 2701–2721.
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.
Morganti, R., N. E. B. Killeen, and C. N. Tadhunter (Aug. 1993). "The radio structures of southern 2-Jy radio sources." In: 263, pp. 1023–1048.
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.
Rudin, Walter (1990). "The Structure of Locally Compact Abelian Groups". In: *Fourier Analysis on Groups*. John Wiley Sons, Ltd. Chap. 2, pp. 35–57.
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.
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.
Williams, C.K.I. and D. Barber (1998). "Bayesian classification with Gaussian processes". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.12, pp. 1342–1351.