Intrinsic Dimension Estimation for Radio Galaxy Zoo using Diffusion Models

Joan Font-Quer Roset
Department of Computer Science
The University of Manchester
Manchester, United Kingdom
joan.font-quer@postgrad.manchester.ac.uk

Devina Mohan
Department of Physics and Astronomy
The University of Manchester
Manchester, United Kingdom
devina.mohan@postgrad.manchester.ac.uk

Anna Scaife*
Department of Physics and Astronomy
The University of Manchester
Manchester, United Kingdom
anna.scaife@manchester.ac.uk

Abstract

In this work, we estimate the intrinsic dimension (iD) of the Radio Galaxy Zoo (RGZ) dataset using a score-based diffusion model. We examine how the iD estimates vary as a function of Bayesian neural network (BNN) energy scores, which measure how similar the radio sources are to the MiraBest subset of the RGZ dataset. We find that out-of-distribution sources exhibit higher iD values, and that the overall iD for RGZ exceeds those typically reported for natural image datasets. Furthermore, we analyse how iD varies across Fanaroff-Riley (FR) morphological classes and as a function of the signal-to-noise ratio (SNR). While no relationship is found between FR I and FR II classes, a weak trend toward higher SNR at lower iD. Future work using the RGZ dataset could make use of the relationship between iD and energy scores to quantitatively study and improve the representations learned by various self-supervised learning algorithms.

1 Introduction

The field of astronomy has experienced a deluge of data from large-scale surveys from new observational facilities such as Euclid [1] and the James Webb Space Telescope [JWST; 2]. This has led to the increased adoption of deep learning models to classify or analyse data [3]. However, these models are constrained by the scarcity of labelled data. This has motivated the development of techniques that can analyse large unlabelled datasets, such as the Radio Galaxy Zoo [RGZ; 4].

According to the manifold hypothesis, high-dimensional data often lies in low dimensional submanifolds [5], whose intrinsic dimensions (iD) reflects the effective number of degrees of freedom in a data distribution. Estimating iD then provides a useful tool to not only analyse data, but also to analyse models trained on the data.

Recent work has shown that diffusion models encode the iD of data manifolds [6] through their score functions. We use this method to estimate the intrinsic dimension of the RGZ dataset using a score-based diffusion model, and compare the resulting estimates against traditional statistical methods (LPCA [7][8], PPCA [9], and MLE [10]).

*The Alan Turing Institute, 96 Euston Rd, London, UK a.scaife@turing.ac.uk

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

We relate the iD estimates to Bayesian neural network (BNN) energy scores obtained from a Hamiltonian Monte Carlo BNN trained on the MiraBest dataset, which is a labelled subset of the RGZ data. These energy scores provide a measure of how in or out of distribution a radio source is relative to the labelled benchmark [11]. Furthermore, we examine how the estimated iD differs across Fanaroff-Riley (FR) morphological classes, using existing labels from recent work on foundation model based radio galaxy classification [12]. The FR classification divides radio galaxies into two morphological types: FR I sources show peak brightness near the core and fade with distance, FR II sources exhibit bright outer lobes separated by a fainter core [13].

2 Estimating intrinsic dimension of data with diffusion models

In a diffusion model, the data distribution p_0 is perturbed using a series of noise-corrupted distributions p_t using a stochastic differential equation [14]. Diffusion models are trained to approximate the score vector $\nabla_{\mathbf{x}} \ln(p_t \mathbf{x})$ with a neural network $s_\theta(x_t, t)$. Once the score function is approximated, it can be used to generate new samples from the data distribution. The recently proposed method provides a theoretical foundation, showing that the approximate local dimensionality of the data manifold can be recovered from the score vector [6]. In experiments with the MNIST dataset, it is found that digits with different geometric complexity have different iDs. For example the digit 1 has iD $= 66$ and the digit 9 has iD $= 152$. Previous work by [15] also estimated the iD of several image datasets using a maximum likelihood estimation (MLE) based method, but later work showed that this severely underestimates iD.

2.1 Theoretical Foundation

Algorithm 1 Estimate the Intrinsic Dimension at $\mathbf{x}_0$ [6]

Require: s_θ – trained diffusion model (score),
t_0 – sampling time,
K – number of score vectors.
Sample $\mathbf{x}_0 \sim p_0(\mathbf{x})$ from the dataset
$d \leftarrow \dim(\mathbf{x}_0)$
$S \leftarrow$ empty matrix
for $i = 1, \ldots, K$ **do**
  Sample $\mathbf{x}_{t_0}^{(i)} \sim \mathcal{N}(\mathbf{x}_0, \sigma_{t_0}^2 I)$
  Append $s_\theta(\mathbf{x}_{t_0}^{(i)}, t_0)$ as a new column to S
end for
$(s_i)_{i=1}^d, (\mathbf{v}_i)_{i=1}^d, (\mathbf{w}_i)_{i=1}^d \leftarrow \mathrm{SVD}(S)$
$\hat{k}(\mathbf{x}_0) \leftarrow d - \arg\max_{i=1,\ldots,d-1}(s_i - s_{i+1})$
return $\hat{k}(\mathbf{x}_0)$ $(s_i)_{i=1}^d, (\mathbf{v}_i)_{i=1}^d, (\mathbf{w}_i)_{i=1}^d$ denote singular values, left and right singular vectors respectively.

Theorem 2.1 *For any point $\mathbf{x} \in \mathbb{R}^d$ sufficiently close to a compact embedded manifold $\mathcal{M}$, and a sufficiently small diffusion time t, the score vector $\nabla_{\mathbf{x}} \ln(p_t(\mathbf{x}))$ points directly at the projection of $\mathbf{x}$ on the manifold.*

Corollary 2.1.1 *The ratio of the projection of the score vector $\nabla_{\mathbf{x}} \ln(p_t \mathbf{x})$ on the tangent space of the manifold $T_{\pi(\mathbf{x})\mathcal{M}}$ to the normal space of the manifold $\mathcal{N}_{\pi(\mathbf{x})\mathcal{M}}$ approaches 0 as t approaches 0.*

Theorem 2.1 only works under certain assumptions: (i) The data lies on a compact, smooth, embedded manifold; (ii) p_0 has a smooth (differentiable) density function defined over the manifold; and (iii) the density function is strictly positive on the manifold (i.e. there are no zero-probability regions). Theorem 2.1 and Corollary 2.1.1 underpin the method's central idea: the score vectors of perturbed points near the manifold will concentrate on the normal space to the manifold at $\mathbf{x}_0$ as $t \to 0$. Therefore, its projection to tangent space vanishes as $t \to 0$, and so the number of vanishing singular values corresponds to the dimension of the tangent space. The iD is then the ambient dimension minus the rank of the normal space. Algorithm 1 describes the method proposed in the paper for estimating the iD at a point $\mathbf{x}_0$ using a trained diffusion model. For each $i \in \{1, \ldots K\}$, a sample

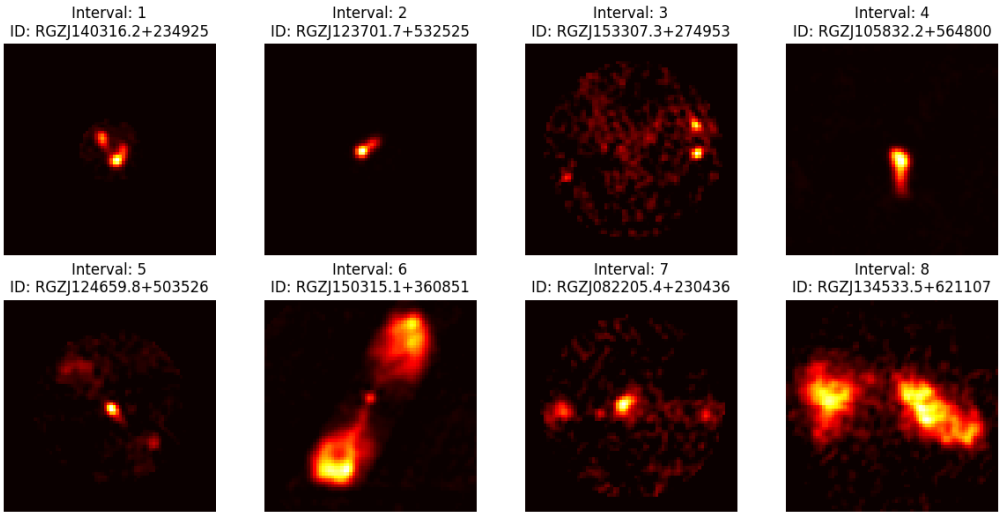


Figure 1: Example images from the Radio Galaxy Zoo dataset taken from different intervals of the mean energy distribution.

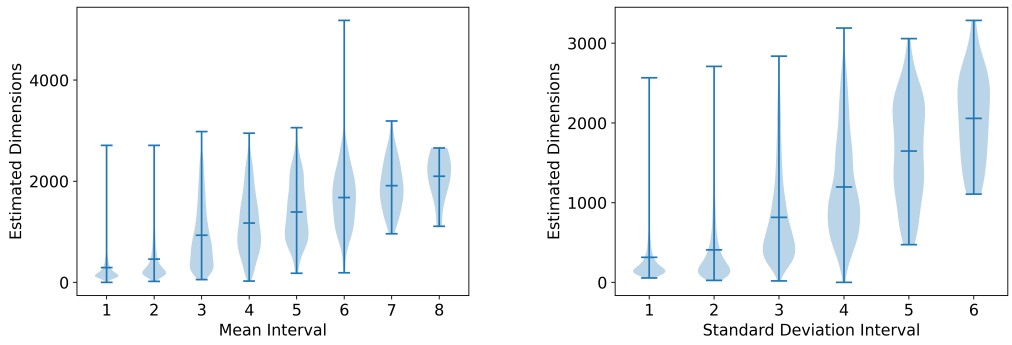
$\mathbf{x}_{t_0}^{(i)}$ is drawn from the forward process transition kernel $p_0(\mathbf{x}_{t_0}^{(i)}|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_0, \sigma_{t_0}^2 I)$, and its score vector $s_\theta(\mathbf{x}_{t_0}^{(i)})$ is computed. Each of these score vectors forms a column of a $d \times K$ matrix S, where d is the ambient dimension of $\mathbf{x}_0$. Singular value decomposition (SVD) is applied to the matrix S, and the iD is estimated as $d - i^*$, where i^* corresponds to the index of the largest gap between singular values $\arg\max_{i=1,\ldots,d-1}(s_i - s_{i+1})$.

3 Experimental Setup

Data We use the Radio Galaxy Zoo Data Release 1 [4] as our unlabelled dataset of radio sources to train the diffusion model. Details of the MiraBest dataset [16] which is used to train the BNN is provided in Appendix A.

Radio Galaxy Zoo Data Release 1 The Radio Galaxy Zoo (RGZ) is a citizen science project which contains extended radio sources from the FIRST and ATLAS radio surveys cross-matched with their host galaxies in the infrared from the WISE and SpitzerSpace Telescope, respectively [4]. Instead of classifying morphologies based on the FR scheme, RGZ classifies sources into the number of components, which are discrete patches of emission enclosed by a contour of constant brightness, and the number of peaks of maximum brightness within a component. The catalogue provides $100,185$ classifications along with a weighted user consensus level. The user consensus level is calibrated by comparing each individual users' classifications to that of an expert for a smaller sample of 20 sources. We use RGZ Data Release 1 in this work which contains classifications with a user consensus level greater than or equal to 0.65. FIRST-based catalogues make up 99.4% of DR1. We select radio sources with angular size > 20 arcsecond. The images are of size 72x72.

Bayesian neural network To impose some structure in our analysis of the iD estimates of RGZ, we create subsets of the dataset using a BNN. We use the BNN based on Hamiltonian Monte Carlo (HMC) from the benchmark provided for radio galaxy classification in [11]. This BNN is trained using the MiraBest dataset to classify radio galaxies into FRI and FRII and performs the best in terms of being able to clearly detect distribution shifted radio galaxies based on energy scores [17]. We pass the RGZ dataset through the trained BNN and obtain a distribution of the scalar energy scores, $\mathrm{E}(x; f) := -T.\log\sum_i^K e^{f_i(x)/T}$, for all the sources in the dataset, x, using the logit values, $f_i(x)$, for each class, i. The temperature term, T, is set to 1. We use $N = 200$ posterior samples from HMC to construct the energy score distributions and calculate the mean and standard deviation of the distribution for each RGZ source. We then fit a log-normal distribution to the distribution of the mean and standard deviation estimates of the entire RGZ dataset and calculate an interval label for each source based on how many standard deviations away a source is from the mean of the mean



(a) (b)

Figure 2: Intrinsic dimension estimates for the RGZ dataset as a function of the (a) mean interval and (b) standard deviation interval of the energy distribution from the Hamiltonian Monte Carlo based Bayesian neural network trained on the MiraBest dataset.

Table 1: Standard deviation interval estimates across benchmarks. "N/A" indicates the method was not applicable due to insufficient samples.

Method	1	2	3	4	5	6
MLE ($m = 5$)	9.318	9.172	17.546	19.157	18.394	20.681
MLE ($m = 20$)	8.182	7.954	15.667	17.207	18.277	20.353
Local PCA	10.571	11.892	16.907	19.144	20.586	21.556
PPCA	4977	3777	N/A	N/A	N/A	N/A

energy values and the mean of the standard deviation values of the energy distributions. We find that the RGZ sources fall into 8 mean intervals and 6 standard deviation intervals. Examples of galaxies from each mean and standard deviation interval are shown in Figure 1 and Figure 4, respectively.

Diffusion Intrinsic Dimension Estimation

We train a score-based diffusion model on the RGZ data using the weighted denoising score matching objective [14] and a Denoising diffusion probabilistic model (DDPM) architecture [18]. The diffusion model is trained for 400 epochs, selecting the model with the lowest validation loss.

We use the method described in Section 2 to estimate the intrinsic dimension of data from different intervals of the mean and standard deviation of energy values. We perturb all the samples $k = 328$ times to calculate the score vector for each noisy point using the diffusion model. The score vectors are collected into a matrix and SVD is performed to estimate iD. The score spectrum plots are shown in Figure 5.

Baseline Intrinsic Dimension Estimation

We estimate the iD for each sample in the RGZ dataset using the MLE method with two neighbouring settings: $m = 5$ and $m = 20$. In addition, we compute the per-sample estimates using local PCA via the Fukunaga-Olsen method [19], setting alpha to 0.05. Probabilistic PCA produces only global estimates, so iD values were instead computed per label.

4 Results & Conclusion

Figure 2 and Tables 2-1 show that the estimated iD increase with interval number for both the mean and standard deviations. This suggests that the object identified with the BNN as out-of-distribution have a higher iD than the objects that are in distribution.

Comparisons with the classical estimators (MLE, LPCA, and PPCA) show that, while they all broadly follow the same trend, the diffusion estimates are much larger in magnitude. This could be due to the

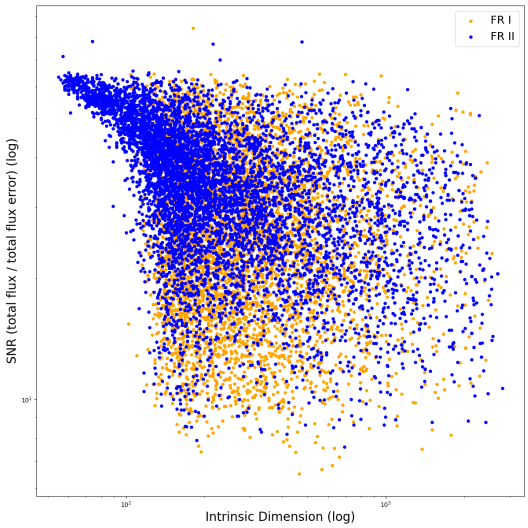


Figure 3: Signal to noise ratio versus the intrinsic dimension for a subset of galaxies labelled by Fanaroff-Riley class. Both axes are shown in logarithmic scale.

Table 2: Mean interval estimates across benchmarks. "N/A" indicates the method was not applicable due to insufficient samples.

Method	1	2	3	4	5	6	7	8
MLE ($m = 5$)	9.249	9.927	12.261	13.437	17.958	21.655	26.092	40.034
MLE ($m = 20$)	8.086	8.645	11.309	12.728	16.174	21.155	25.298	38.033
Local PCA	10.702	12.064	15.368	16.653	17.331	17.523	16.725	17.889
PPCA	4357	3638	N/A	N/A	N/A	N/A	N/A	N/A

fact that the classical methods tend to underestimate the iD of high dimensional data [20]. In addition, the diffusion iD estimates are also much larger compared to those reported for natural image datasets in previous work [15, 6]. This could be due to the inherently noisy nature of radio astronomy data.

Figure 3 shows the relationship between the signal-to-noise ratio (SNR) and the iD of each galaxy, along with their FR morphology. Overall, there is no strong relationship between SNR and iD. However, at low iD values (< 100), FR II sources tend to exhibit higher SNRs, suggesting that galaxies with lower iD are less affected by noise. On the other hand, if a galaxy has a high SNR, it does not necessarily have a low iD.

In future work, we will extend this approach to estimate the iD using the features learned by self-supervised learning based models trained on the RGZ dataset ([21]). We will use the results to determine what degree of compression can be achieved for different subsets of the data.

Acknowledgments and Disclosure of Funding

This work was supported by the Engineering and Physical Sciences Research Council EP/Y030826/1. AMS gratefully acknowledges support from an Alan Turing Institute AI Fellowship EP/V030302/1. This work was done under the supervision of Anna Scaife, Julia Handl, and Mingfei Sun.

References

[1] Y Mellier, Abdurroúf Abdurroúf, JA Acevedo Barroso, et al. Euclid. i. overview of the euclid mission. *Astronomy & Astrophysics*, 2024.

[2] Klaus M Pontoppidan, Jaclyn Barrientes, Claire Blome, et al. The jwst early release observations. *The Astrophysical Journal Letters*, 936(1):L14, 2022.

[3] Natalie EP Lines, Tian Li, Thomas E Collett, et al. The revolution in strong lensing discoveries from euclid. *Nature Astronomy*, 9(8):1116–1122, 2025.

[4] O Ivy Wong, L Rudnick, H Andernach, et al. Radio galaxy zoo data release 1: 100185 radio source classifications from the first and atlas surveys. *Monthly Notices of the Royal Astronomical Society*, 536(4):3488–3506, 2025.

[5] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[6] Jan Pawel Stanczuk, Georgios Batzolis, Teo Deveney, and Carola-Bibiane Schönlieb. Diffusion models encode the intrinsic dimension of data manifolds. In *Forty-first International Conference on Machine Learning*, 2024.

[7] K. Fukunaga and D.R. Olsen. An algorithm for finding intrinsic dimensionality of data. *IEEE Transactions on Computers*, C-20(2):176–183, 1971.

[8] Mingyu Fan, Nannan Gu, Hong Qiao, and Bo Zhang. Intrinsic dimension estimation of data by principal component analysis, 2010.

[9] Thomas P. Minka. Automatic choice of dimensionality for pca. In *Proceedings of the 14th International Conference on Neural Information Processing Systems*, NIPS'00, page 577–583, Cambridge, MA, USA, 2000. MIT Press.

[10] Elizaveta Levina and Peter J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, NIPS'04, page 777–784, Cambridge, MA, USA, 2004. MIT Press.

[11] Devina Mohan and Anna M. M. Scaife. Evaluating bayesian deep learning for radio galaxy classification. In Negar Kiyavash and Joris M. Mooij, editors, *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pages 2587–2597. PMLR, 15–19 Jul 2024.

[12] Nutthawara Buatthaisong, Inigo Val Slijepcevic, Anna M. M. Scaife, et al. Radio Galaxy Zoo: Morphological classification by Fanaroff-Riley designation using self-supervised pre-training. *arXiv e-prints*, page arXiv:2509.11988, September 2025.

[13] B. L. Fanaroff and J. M. Riley. The morphology of extragalactic radio sources of high and low luminosity. *mnras*, 167:31P–36P, May 1974.

[14] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, et al. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[15] Phil Pope, Chen Zhu, Ahmed Abdelkader, et al. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021.

[16] Fiona Alice May Porter. MiraBest Batched Dataset, 11 2020.

[17] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[19] K. Fukunaga and D. R. Olsen. An algorithm for finding intrinsic dimensionality of data. *IEEE Trans. Comput.*, 20(2):176–183, February 1971.

[20] Haiquan Qiu, Youlong Yang, and Hua Pan. Underestimation modification for intrinsic dimension estimation. *Pattern Recognition*, 140:109580, 2023.

[21] Inigo V Slijepcevic, Anna MM Scaife, Mike Walmsley, et al. Radio galaxy zoo: towards building the first multipurpose foundation model for radio astronomy with self-supervised learning. *RAS Techniques and Instruments*, 3(1):19–32, 2024.

[22] Fiona AM Porter and Anna MM Scaife. Mirabest: a data set of morphologically classified radio galaxies for machine learning. *RAS Techniques and Instruments*, 2(1):293–306, 2023.

[23] H. Miraghaei and P. N. Best. The nuclear properties and extended morphologies of powerful radio galaxies: The roles of host galaxy and environment. *Monthly Notices of the Royal Astronomical Society*, 466(4):4346–4363, 5 2017.

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

[25] Kevork N. Abazajian, Jennifer K. Adelman-McCarthy, Marcel A. Agüeros, et al. The Seventh Data Release of the Sloan Digital Sky Survey. *apjs*, 182(2):543–558, June 2009.

[26] J. J. Condon, W. D. Cotton, E. W. Greisen, et al. The NRAO VLA Sky Survey. *aj*, 115(5):1693–1716, May 1998.

[27] Robert H. Becker, Richard L. White, and David J. Helfand. The FIRST Survey: Faint Images of the Radio Sky at Twenty Centimeters. *apj*, 450:559, September 1995.

[28] Bernard L Fanaroff and Julia M Riley. The morphology of extragalactic radio sources of high and low luminosity. *Monthly Notices of the Royal Astronomical Society*, 167(1):31P–36P, 1974.

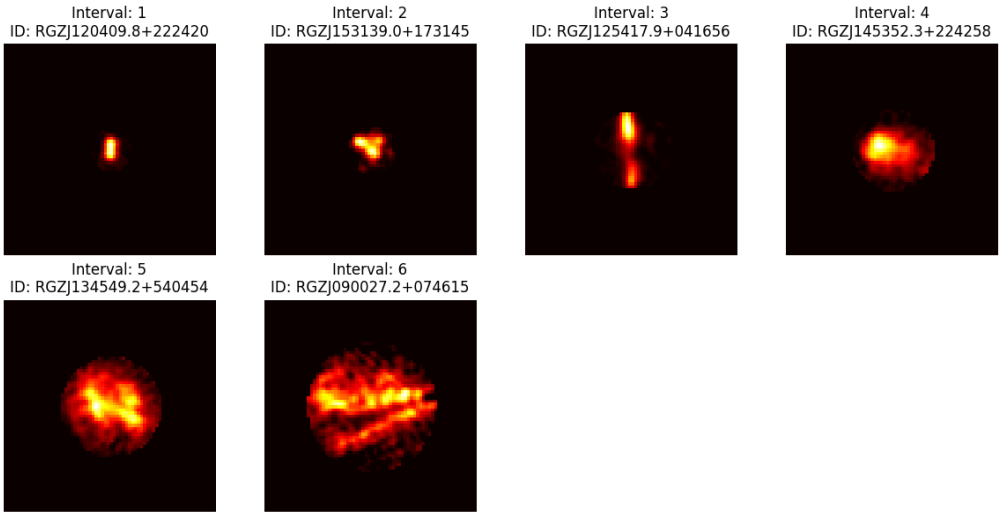


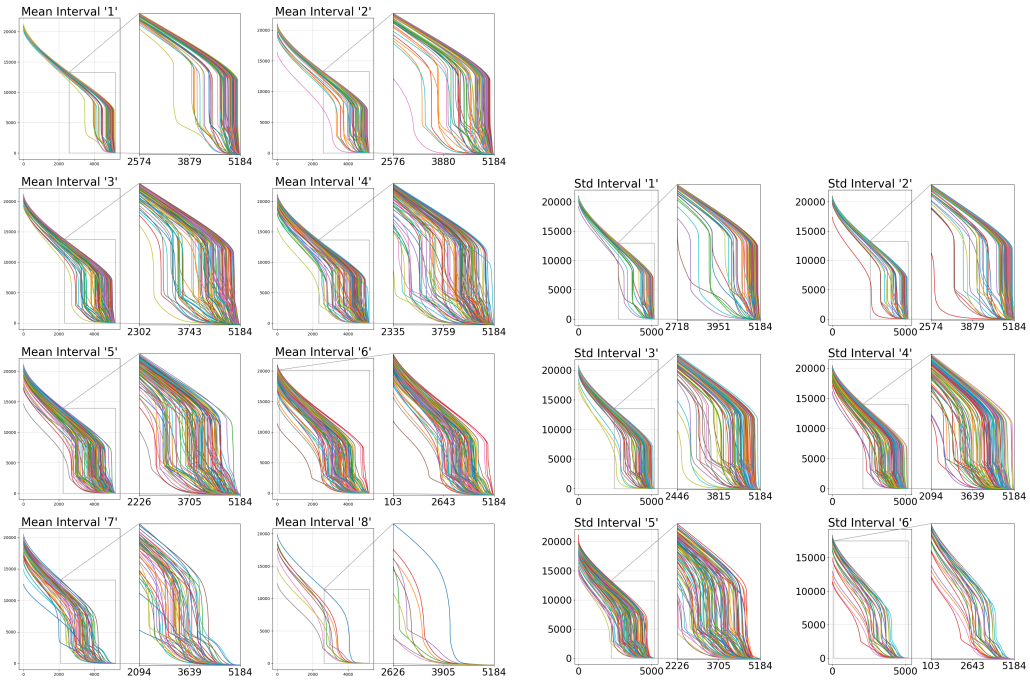
Figure 4: Images from the Radio Galaxy Zoo dataset from different intervals of the standard deviation of energy scores.

A MiraBest dataset

MiraBest Dataset The MiraBest dataset used in this work consists of 1256 images of radio galaxies of 150×150 pixels pre-processed to be used specifically for deep learning tasks [22]. The dataset was constructed using the sample selection and classification described in [23], who made use of the parent galaxy sample from [24]. Optical data from data release 7 of Sloan Digital Sky Survey [SDSS DR7; 25] was cross-matched with NRAO VLA Sky Survey [NVSS; 26] and Faint Images of the Radio Sky at Twenty-Centimeters [FIRST; 27] radio surveys. The galaxies are labelled using the FRI and FRII morphological types based on the definition of [28] and further divided into their subtypes. In addition to labelling the sources as FRI, FRII and their subtypes, each source is also flagged as 'Confident' or 'Uncertain' to indicate the human classifiers' confidence while labelling the dataset. In this work we use the MiraBest Confident subset and consider only the binary FRI/FRII classification. The training and validation sets are created by splitting the predefined training data into a ratio of 80:20. The final split consists of 584 training samples, 145 validation samples, and 104 withheld test samples.

B Score Spectra

Figure 5 shows the score spectra for 100 randomly selected samples for each label.



(a) (b)

Figure 5: Score spectrum plots for up to 100 randomly selected RGZ sources from different (a) mean and (b) standard deviation intervals of the energy distribution. Here, the x-axis shows the singular values which go up to the ambient dimension (72x72 for RGZ images) and the y-axis shows the magnitude of each singular value. The point at which there is a sharp drop in singular values indicates the normal dimension. This can be subtracted from the ambient dimension to calculate the intrinsic dimension.