

# Generative Adversarial Networks for Galaxy Classification Using Convolutional Neural Networks

Polina Petrova

ppetrova@umass.edu

Aryan Singh

arysingh@umass.edu

## Abstract

*The classification of galaxies poses significant challenges due to inherent data imbalances among various morphological classes. This research addresses the pressing issue of accurately classifying galaxies using Convolutional Neural Networks (CNNs), where traditional approaches often struggle with underrepresented categories. To mitigate this problem, we propose a novel combined model utilising a Conditional Generative Adversarial Network (cGAN) to generate synthetic images that augment the Galaxy Zoo 2 dataset, which contains approximately 250,000 labelled galaxy images. Our methodology begins with enhancing image resolution using Super Resolution GANs (SRGAN), followed by training the cGAN to produce additional samples for each class, thereby addressing class imbalance without resorting to deeper networks. We draw on foundational works, including studies on deep generative models for galaxy image simulations and prior CNN applications in galaxy classification, to inform our approach. Our model's performance will be rigorously evaluated, comparing results from training on both real and synthetic data against traditional CNNs trained on the original dataset. Through this work, we aim to contribute to the ongoing efforts to improve automated galaxy classification and provide a scalable solution to the prevalent issue of data imbalance in astronomical research.*

## 1. Introduction

With rapid development in technology comes a phenomenon that is referred to as the *data deluge*, – a consequence of an overwhelming influx of data and insufficient resources to process it. Astronomers, in particular, contend with this challenge; estimates suggest the observable universe holds between 100 billion and 200 billion galaxies [3]. A single photograph of a small sky section can capture up to 25,000 galaxies, and the resulting daily data volume overwhelms the limited pool of experts available to classify them [9]. To address this challenge, astrophysicists launched the *Galaxy Zoo* project in 2007, inviting citizen

scientists to help classify over 900,000 galaxies, marking a transformative moment in data processing through public participation [8].

Following the success of the *Galaxy Zoo*, advancements in machine learning spurred efforts to automate galaxy classification. Modern approaches employ Convolutional Neural Networks (CNNs) to recognise patterns with minimal human input. However, a persistent challenge is the quality and distribution of available data. While the *Galaxy Zoo* project provided a substantial dataset, imbalances in class representation can skew training, causing models to favour more frequent classes. Our research specifically addresses this class imbalance in the Galaxy Zoo 2 dataset – a collection of categorised images taken from the Sloan Digital Sky Survey (SDSS) – where certain galaxy types are underrepresented. Building deeper networks is a common workaround to address this issue, but we argue that this approach only sidesteps the core problem.

Ideally, a large, balanced dataset would improve model accuracy, but limitations in space imaging and classification complexity make this difficult. To tackle this, we propose a novel solution using Generative Adversarial Networks (GANs) to generate synthetic images for underrepresented galaxy classes in the Galaxy Zoo 2 dataset, which we then use to train a CNN. We expect that by augmenting our data with synthetic images, we can improve classification accuracy across all classes. Our evaluation will compare the GAN-CNN model's performance against traditional CNNs trained on imbalanced data, particularly examining accuracy gains in classifying underrepresented classes. With this combined GAN-CNN model, we aim to address class imbalance directly, eliminating the need for deeper networks as a compensatory measure.

## 2. Related work

Lahav et al. [5] offers one of the first discussions of using neural networks in the galaxy classification problem. The study clarifies the role of Artificial Neural Networks (ANNs) in galaxy classification by demonstrating their ability to replicate human classification using ESO-LV galaxy data. ANNs achieve comparable accuracy to human ex-

perts, operating within 2 T-type units. The authors argue that ANNs provide a robust statistical framework, improving on linear methods through their capacity for non-linear modelling. While the paper does not cover all classification methods, it emphasises the potential of unsupervised algorithms to discover new features in galaxy data without external guidance. It also highlights the importance of integrating dynamic properties and multiwavelength data to enhance the classification process, as we now have in the Galaxy Zoo 2 dataset. This study lays the groundwork for our exploration into data-driven galaxy classification.

Fussell and Moews [2] demonstrate the effectiveness of using GANs to augment limited datasets of galaxy images. The authors find that the original DCGAN architecture can generate realistic galaxy images that align closely with real galaxy data when statistically evaluated. To achieve higher-resolution synthetic galaxies, they introduce a chained approach using StackGAN as a second stage, which overcomes DCGAN’s limitations at higher resolutions. By evaluating physical property distributions of the generated galaxies and confirming their similarity to real data, the study suggests these synthetic images can effectively expand real galaxy datasets. This augmentation is beneficial for various tasks, including galaxy classification, segmentation, deblending, and calibration of shape measurement algorithms. Ultimately, this research highlights GAN architectures as valuable resources for astronomy, providing scalable data for deep learning models that require extensive training samples. We will adopt a similar approach, benchmarking our GAN-generated data against real data in our model pipeline before feeding them into our CNN.

Kim and Brunner [4] address the limitations and potential improvements for using CNNs in galaxy classification, emphasising concerns about overfitting due to limited training data. The authors note that while CNNs have shown promise, their model did not significantly outperform traditional machine learning models relying on summary catalog data, likely due to data constraints. Collecting additional spectroscopic training images could mitigate overfitting and enhance CNN performance. However, they argued that the process is costly and time-intensive. For future work, the authors suggest training multiple network architectures and combining them, a strategy that has proven effective in other galaxy classification challenges. Integrating CNN with other classifiers in a hybrid model could also yield improvements, as demonstrated in past studies where blending diverse classification approaches outperformed any single method. We will directly address these concerns in our research, providing a hybrid model that will even out and expand our training data to mitigate overfitting.

Walmsley et al. [1] focus on the limitations of deep learning for galaxy morphology, which often overlook uncertainty in labelling. They introduce a Bayesian CNN model

to capture probabilistic predictions for galaxy morphology, leveraging sparse Galaxy Zoo labels. Using Monte Carlo Dropout and active learning, their model selects informative galaxies for labelling, enhancing classification accuracy with fewer labels. This approach, essential for large-scale surveys, offers insights into morphology and astrophysical connections. We aim to incorporate similar probabilistic and active learning strategies in future iterations of our model for effective scaling.

## 3. Method

### 3.1. Galaxy Zoo 2 Dataset

**RECOMMENDATION FROM PROFESSOR:** Consider adding a section introducing the Galaxy Zoo 2 dataset. Detail the various classes, the image count for each (including some sample images if possible), and discuss any class imbalances. Outline which classes you intend to expand and the scale of this expansion. Additionally, provide an overview of the dataset post-augmentation to show the changes in its composition.

The Galaxy Zoo 2 dataset contains detailed galaxy morphology data. The Galaxy Zoo 2 dataset contains a total of 37 morphological galaxy classes. There is a severe imbalance in the dataset, with the most frequent classes containing over 40,000 instances and the least frequent classes containing under 3 instances. This posed a serious issue, as there was no way to effectively train the CNN model of the most underrepresented classes.

*Initial Data and Image Preprocessing:* We first cleaned the original dataset, keeping only columns containing debiased data, as well as Object ID, right ascension value and declination value for each galaxy image. The debiased columns indicate the debiased fraction of votes in favour of each decision tree response for each task. Debiased: which is the weighted vote fraction adjusted for classification bias

### 3.2. Image Enhancement and Generation Using GAN

*Image Enhancement:* We start with the Galaxy Zoo 2 dataset, which contains a substantial but imbalanced set of labelled galaxy images. Since GANs can benefit from high-quality images, we will first enhance image quality by implementing a Super-Resolution Generative Adversarial Network (SRGAN). The SRGAN will upscale images, improving the resolution to better capture fine details crucial for accurate classification. We will tune SRGAN parameters, including the number of generator and discriminator layers, learning rate and regularisation parameters, to achieve optimal enhancement without introducing artifacts.

*Synthetic Data Generation Using cGAN:* To address the data imbalance across galaxy classes, we will implement a Conditional Generative Adversarial Network (cGAN) to

generate additional samples for each class. The cGAN will be conditioned on class labels, ensuring that synthetic images resemble galaxies of specific morphological types. By generating images conditioned on each class, we aim to balance the dataset, allowing the CNN to learn from a more uniform distribution of galaxy characteristics. Hyperparameters, such as the batch size, learning rate and the trade-off between generator and discriminator objectives will be carefully selected to ensure diverse image generation.

**GAN-CNN Pipeline Integration:** The generated synthetic images, alongside the SRGAN-enhanced real images, will be used to train the CNN model. The CNN architecture will feature several convolutional and pooling layers to extract relevant features from galaxy images, followed by fully connected layers for classification. Each layer configuration, activation function, and regularisation strategy will be optimised specifically for galaxy image data. Initially, the CNN will be trained solely on the original dataset to establish a baseline performance. Then, we will re-train the CNN using a combination of real and synthetic data, assessing improvements in classification accuracy and robustness across all classes.

**Evaluation:** We will use accuracy, precision, recall, and F1-score as quantitative metrics to evaluate the CNN’s classification performance. Additionally, we will compare results from the CNN trained on original images alone against those trained with the augmented dataset, quantifying improvements attributable to our image generation. Visualisations, such as ROC curves, will provide further insights into class-wise performance, highlighting the impact of the GAN-CNN pipeline on addressing class imbalances.

## 4. Results

**RECOMMENDATION FROM PROFESSOR:** Consider adding some images to show the generated image quality.

The SRGAN model has been trained to upscale images by a factor of 2x. Visual inspection of the upscaled images shows noticeable improvements in clarity and detail, especially in smaller galaxy structures. We achieved an average PSNR value of 24.5 across the upscaled images, indicating a high degree of similarity between the generated high-resolution images and their lower-resolution counterparts.

## 5. Conclusion

## References

- [1] M. Walmsley et al. Galaxy zoo: probabilistic morphology through bayesian cnns and active learning. *Monthly Notices of the Royal Astronomical Society*, 491(2):1554–1574, 2019. [2](#)
- [2] L. Fussell and B. Moews. Forging new worlds: high-resolution synthetic galaxies with chained generative adversarial networks. *Monthly Notices of the Royal Astronomical Society*, 485(3):3203–3214, 2019. [2](#)
- [3] E. Howell. How many galaxies are there?, 2018. Accessed: Oct. 30, 2024. [1](#)
- [4] E. J. Kim and R. J. Brunner. Star–galaxy classification using deep convolutional neural networks. *Monthly Notices of the Royal Astronomical Society*, 464(4):4463–4475, 2016. [2](#)
- [5] O. Lahav, A. Nairn, L. Sodré, and M. C. Storrie-Lombardi. Neural computation as a tool for galaxy classification: methods and examples. *Monthly Notices of the Royal Astronomical Society*, 283(1):207–221, 1996. [1](#)
- [6] F. Lanusse, R. Mandelbaum, S. Ravanbakhsh, C.-L. Li, P. Freeman, and B. Póczos. Deep generative models for galaxy image simulations. *Monthly Notices of the Royal Astronomical Society*, 504(4):5543–5555, 2021.
- [7] V. Lukic and M. Brüggen. Galaxy classifications with deep learning. *Proceedings of the International Astronomical Union*, 12(S325):217–220, 2016.
- [8] C. McGourty. Scientists seek galaxy hunt help, 2007. Accessed: Oct. 30, 2024. [1](#)
- [9] E. Sauers. Webb telescope reveals more galaxies in a snapshot than hubble’s deepest survey, 2023. Accessed: Oct. 30, 2024. [1](#)
- [10] K. Sharma, A. Kembhavi, T. Sivarani, S. Abraham, and K. Vaghmare. Application of convolutional neural networks for stellar spectral classification. *Monthly Notices of the Royal Astronomical Society*, 491(2):2280–2300, 2019.