Effective Augmentation Methods for CNN-Based Galaxy Morphology Classification

Min Ki, Hong

July 2025

Abstract

Morphological classification of galaxies has been a foundational practice in astronomy since Edwin Hubble introduced his scheme, as it provides key insights into the physical processes governing galaxy evolution and star formation. However, given the vast number of galaxies in modern surveys, it is no longer feasible for humans to classify them all manually; therefore automated classification has become essential. One widely used approach for automated classification involves the application of machine learning techniques, specifically convolutional neural networks, which are particularly effective for image data.

In this experiment, various data augmentation methods were applied to galaxy images from the Sloan Digital Sky Survey (SDSS), using classifications from the Galaxy Zoo 2 project. The goal was to evaluate how effective each augmentation method was for automated morphological classification using convolutional neural networks. Distributions of test accuracies for each augmentation method were obtained, and the statistical significance of each augmentation method, compared to the baseline model, was assessed using the Mann-Whitney U test. Among the augmentation methods, random rotation, stretching and shearing, and flipping showed statistically significant improvements over the baseline model.

For additional analysis, these three augmentation methods were combined to train the final model. The final model achieved an overall accuracy of 83.28%, with 88.07% accuracy for classifying non-disk galaxies and 79.19% for disk galaxies. However, signs of underfitting during training — despite modifications to the model capacity — indicated that the constraint on achieving higher accuracy stemmed from the quality and ambiguity of the training data itself, rather than from the model capacity.

1 Introduction

Morphological classification of galaxies has been a foundational practice in astronomy since Edwin Hubble introduced his scheme, commonly referred to as Hubble sequence or Hubble's tuning fork [1]. This scheme was later expanded by de Vaucouleurs, whose classification added structural detail to the original sequence, forming the basis of modern galaxy morphology.

Studying galaxy morphology provides key insights into the physical processes governing galaxy evolution and star formation [2]. Moreover, the distribution of the morphological types across cosmic time can be used to examine the cosmological model, making a morphology a relevant parameter in observational cosmology [3].

Given the vast number of galaxies in modern surveys, it is no longer feasible for humans to classify them all manually. This challenge is further amplified by the emergence of next-generation observatories such as the James Webb Space Telescope (JWST) or the Square Kilometre Array (SKA). As a result, automated classification has become essential for preparing, interpreting, and analysing these large datasets. One widely used approach involves the application of machine learning techniques.

Among machine learning techniques, convolutional neural networks (CNNs), a class of neural network models particularly effective for image data, have demonstrated remarkable progress in tasks ranging from object recognition to generative modeling. Astronomy is now one of the key fields benefiting from this progress, particularly in automated image classification task.

The Galaxy Zoo 2 project was a citizen science project and successor to Galaxy Zoo 1. It provided the morphological classification for 304,122 galaxies from the Sloan Digital Sky Survey (SDSS), based on crowd-sourced vote contributed by volunteer citizen scientists [4].

In this experiment, various data augmentation methods were applied to galaxy images from the SDSS, using classifications from the Galaxy Zoo 2 project, to evaluate their effectiveness for the morphological classification. The results were analysed in terms of both performance and morphological interpretability.

2 Theory

2.1 Convolutional Neural Network

A neural network is a machine learning model inspired by biological neurons. Each artificial neuron receives inputs from the previous layer, multiplies them by weights, adds a bias, and applies an activation function to produce an output. The output of j-th neuron in the l-th layer

is defined as

$$a_j^l = f\left(\sum_k w_{jk}^l a_k^{l-1} + b_j^l\right) \tag{1}$$

where a_k^{l-1} is the output from the previous layer, w_{jk}^l are the weights, b_j^l is the bias and $f(x)$ is an activation function.

Neural networks are composed of many such neurons connected across layers. There are various types of neural networks depending on their structure and application, including feedforward neural network, transformers, and convolutional neural networks (CNNs) neural network.

CNNs are specifically designed for processing data with spatial structure such as images, using a

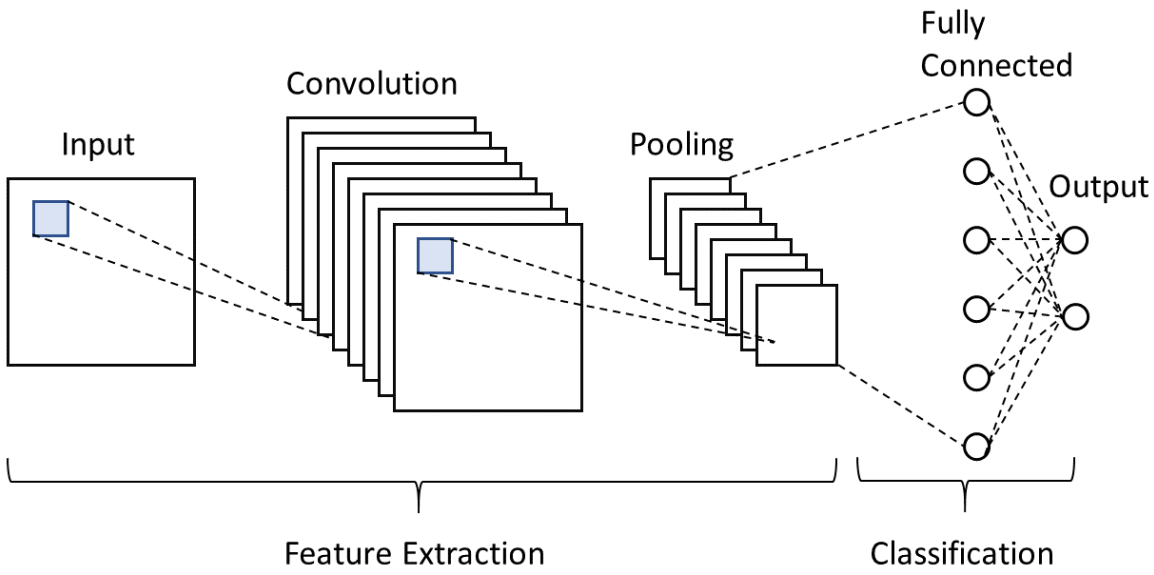

Figure 1: Schematic diagram of basic convolutional network. It shows how initial input image is processed through convolutional layer, pooling layer and finally fully connected layer

concept of convolution, a mathematical operation. The detailed structure of CNNs are depicted in the Figure 1. Filters (also called kernel), which contain the learnable weights and biases, move across the input data and apply the operation described in Equation 1, producing output. The output is known as a feature map, as it captures local patterns and highlights specific visual features such as edges or textures. Due to the limited size of filter, each operation only interacts with a small local region of the input, and this allows CNNs to be particularly effective at learning spatial features and recognising image structures.

After the convolution, the activation function is applied, and the resulting feature maps are

typically passed through a pooling layer. Pooling operation, such as max pooling or average pooling, reduce the resolution of feature maps while preserving the most relevant information. It essentially makes the network more efficient and less sensitive to small translations or distortion.

This sequence from convolution to pooling layer can be repeated multiple times to extract more complex features. Finally, the output is flattened and passed to one or more fully connected layers, where all neurons are fully connected to each other with learnable weights and biases to give a final classification. In terms of supervised image classification, the final output is a vector where each element represents the confidence for a specific class for given input.

The entire system is trained by backpropagation. First, a loss function is defined such that its value decreases as the network's prediction becomes closer to the true labels. Common choices include a mean squared loss or a cross-entropy loss. Since finding an analytical solution for the global minimum of the loss function is generally intractable, the minimisation is typically performed using gradient-based methods. This involves computing a gradient of the loss function with respect to each parameter and adjusting the parameters accordingly to minimise the loss. This process is known as gradient descent, or stochastic gradient descent (SGD) when the gradient is estimated by a small batch of training data. In practice, variants of SGD are used, such as Adaptive Moment Estimation (ADAM), which includes adaptive learning rates and momentum to help convergence.

2.2 Image Augmentation

One of the most important challenges in building neural networks is ensuring that the model generalises well to unseen data. Various regularisation techniques are employed to improve generalisation, including weight decay, dropout, batch normalisation etc. Among these, one of the most effective strategies is to increase the size of training data. However, acquiring additional data is often expensive or impractical. In such cases, data augmentation offers an efficient alternative by artificially expanding the dataset by transforming the original samples. In image classification tasks, this may involve transformations of images such as rotation, flipping, shearing or colour adjustment, that simulate natural variations encountered in real-world scenarios.

A more recent and increasingly popular augmentation method is generating new data using generative AI models. Although these models cannot produce entirely new data that is independent of training set, they can recombine existing patterns to create diverse examples that differ meaningfully from the originals. This process can reveal underlying structure not explicitly visible in the original data and provide richer representations, thereby enhancing generalisation performance.

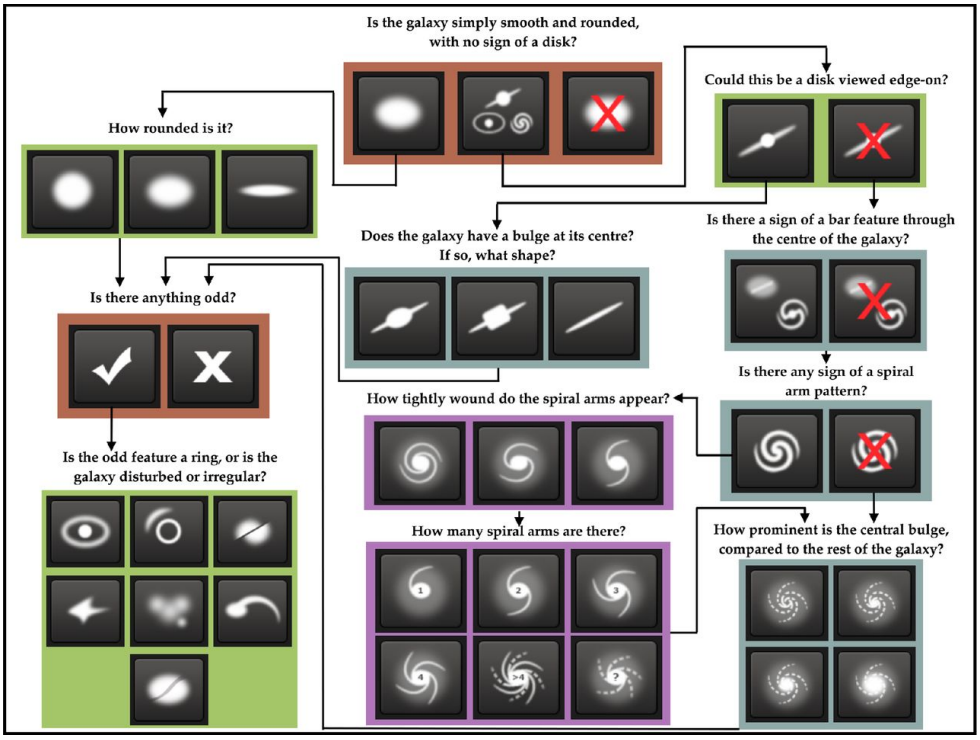

Figure 2: This diagram illustrates the decision tree used in the Galaxy Zoo 2 project to classify galaxy morphologies based on visual features. Adapted from Willett et al. (2013) [4].

3 Methodology

3.1 Data Processing

The images data and their labels were downloaded from Galaxy Zoo 2 website [5]. Each image had a label determined by a decision tree described by Figure 2. Although the Galaxy Zoo 2 dataset provided detailed morphological classifications, incorporating the full hierarchy would have required multiple CNNs for each classification stage or a model with explicit hierarchical structure. Since this was beyond the scope of the experiment, the dataset was categorized only by the primary decision stage, a presence of disk. Additionally, any other odd featured images such as galaxies merging each other, or non-galaxy objects were removed from the dataset as they also do not fall into the same level of categorisation. Figure 3 shows example galaxy images from two morphological categories used in this experiment, non-disk (left) and disk (right). These examples illustrate the typical visual features used in classification. Disk galaxies often exhibit flattened structure with spiral arms. While non-disk galaxies appear more compact or elliptical in structure. Some cases are visually distinct but others show ambiguity.

The original RGB images, each sized 424×424 pixels, were cropped to 256×256 pixels to remove unnecessary margins and reduce computational cost. Subsequently, the images were converted to grayscale and normalised to the range $[-1, 1]$. While RGB colour information, particularly

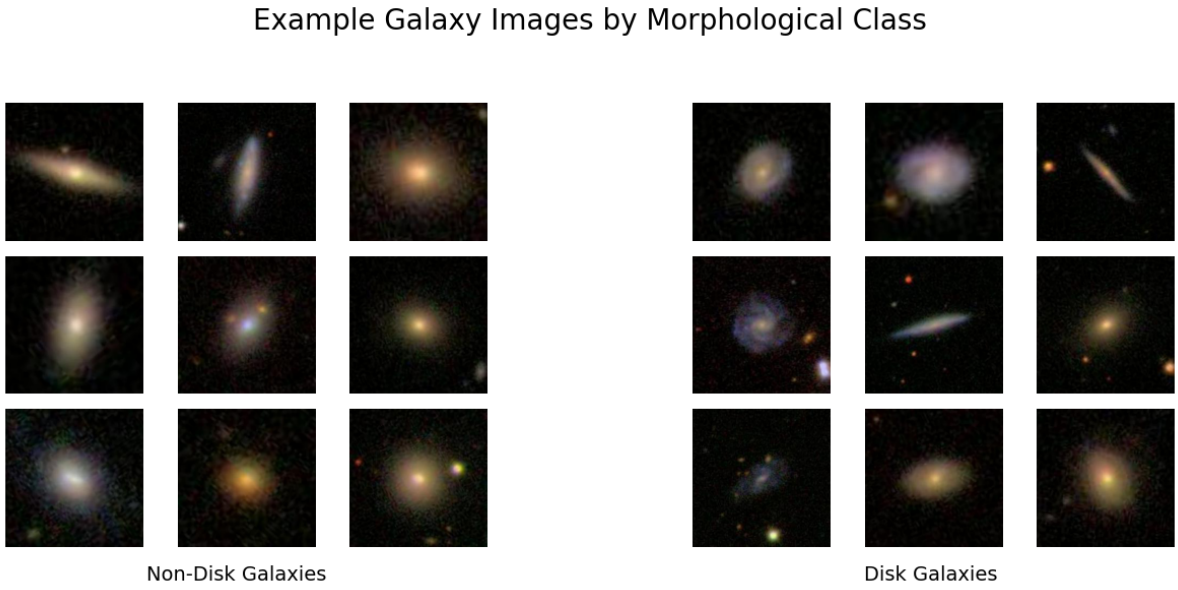

Figure 3: Example galaxy images (256×256 pixels) from the two morphological classes used in this study. The left panel shows non-disk galaxies, while the right panel shows disk galaxies.

when combined with redshift corrections from Galaxy Zoo 2 project, can be valuable in certain contexts, grayscale images were used in this experiment to emphasise morphological features for classification. The final data set consisted of 5438 non-disk galaxies and 6378 disk galaxies, which were randomly split into training, validation, and test sets in a ratio of 8:1:1.

To evaluate the effectiveness of various augmentation methods, each method was applied independently. The first augmentation involved stretching and shearing to mimic the variation in the shape and the orientation of the galaxies. The images were randomly resized by a factor ranging from 0.8 to 1.2, and sheared horizontally and vertically by up to $\pm 10°$. The second augmentation applied random rotations in the range $[0°, 360°)$ to simulate the arbitrary orientation of galaxies. The third involved adding noise in the images. It was done by adding Gaussian noise with standard deviation of 0.01 to each pixel, mimicking observational noise. A fourth was brightness and contrast jitter. Both parameters were randomly applied within $\pm 20\%$ to simulate minor photometric variations. The fifth method was to randomly flip the images horizontally and vertically. The sixth augmentation applied Gaussian blurring on an entire image with a kernel size of 3 to the images with a probability of 20%.

Lastly, a separate augmentation involved generating synthetic images using a generative AI model. This was done using a Stable Diffusion model with Low-Rank Adaptation (LoRA), trained on 256×256-pixel images from each morphological class, using the kohya-ss training framework [6]. A single-word caption, such as disk galaxy or non-disk galaxy, was used for training. Stable Diffusion v1.5 was used for the base model to train LoRA [7]. The training was

done over 10 epochs and the other training hyper-parameters were left at their default values provided in kohya-ss.

Once trained, the LoRA model was used in a text-to-image setting to generate synthetic images from single-word prompts, disk galaxy or non-disk galaxy. A total of images equal in number to the original training set were generated, each at a resolution of 256×256 pixels. Image generation was performed using the Euler sampling method with 20 steps and a classifier-free guidance (CFG) scale of 7. The generated images were then added to the original training set.

All augmentation methods, with the exception of the AI-generated images, were implemented using the Torchvision library.

3.2 Model Training

Through experimental runs on the training on validation sets, the model structure and learning hyperparameters were determined. The final model architecture consisted of seven convolutional layers followed by two fully connected layers. The convolutional layers began with 16 feature maps, doubling at each layer up to 1024 at the seventh convolutional layer. Each convolutional layer was followed by batch normalisation, a Rectified Linear Unit (ReLU) activation function, and max pooling. The output of the final convolutional layer with dimensions of $1024 \times 2 \times 2$ was flattened and passed through a fully connected layer reducing the size to 256, followed by another fully connected layer of size 2, corresponding to the number of classes. Dropout with a rate of 0.5 was applied after each fully connected layer.

The model was trained using the cross-entropy loss function and optimised with Adam optimiser. The learning rate was set to 0.0001, with a weight decay of 0.0001 to provide L2 regularisation. Exponential decay rates for the first the second moment in the Adam optimiser were set to be 0.9 and 0.999, respectively. A learning rate scheduler was employed to improve convergence, reducing the learning rate by a factor of 0.5 when the validation loss did not improve for three epochs. A batch size of 24 was used for the model incorporating AI-generated synthetic images, while a batch size of 12 was used for all other models.

Different loss weights were assigned to each category based on the number of images, with higher penalties applied to categories containing more samples. This was done to prevent the model from becoming biased toward majority classes and focusing only on specific categories.

To detect signs of overfitting, accuracy and loss were continuously monitored on both training and validation sets. Model checkpoints were saved at each epoch, and the best-performing model for each run was selected based on the lowest validation loss. This procedure was repeated multiple times to obtain a distribution of test accuracies for each augmentation method. The statistical

significance of each augmentation methods, compared to the baseline model, was assessed using the Mann-Whitney U test. The number of repetitions varied depending on the level of statistical confidence required to assess significance.

Finally, a model incorporating all augmentation methods that showed significant improvement was trained, and its performance was analysed.

4 Results

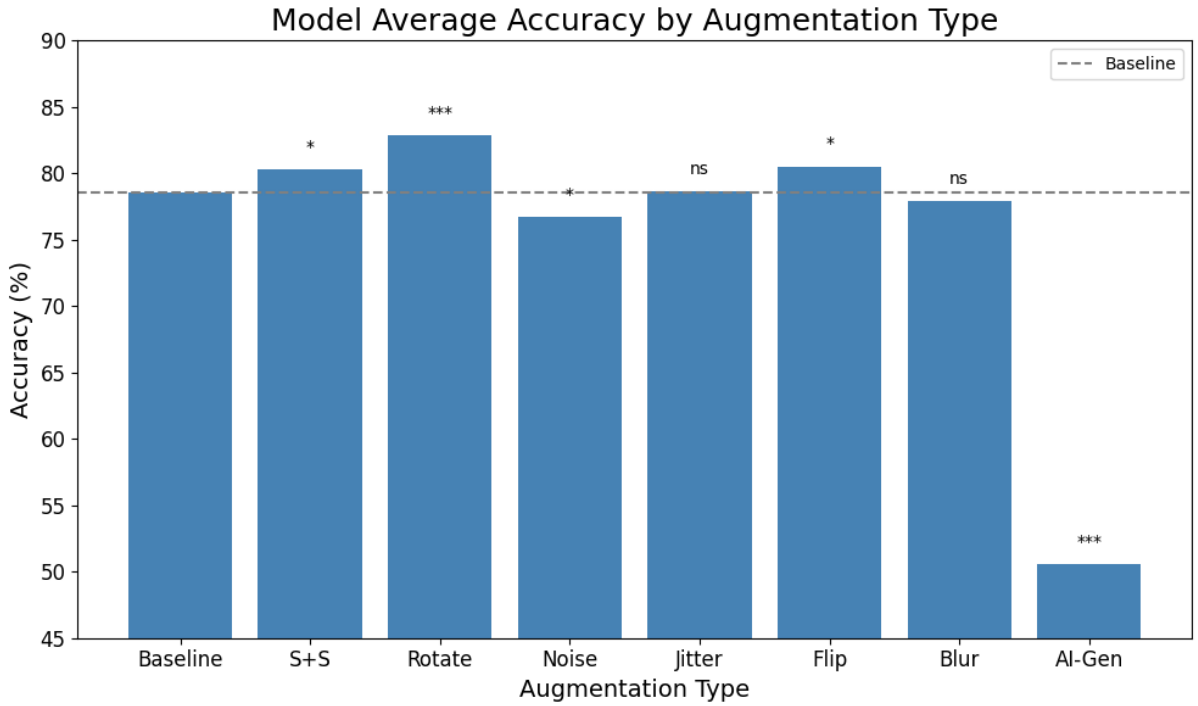

Figure 4: Average classification accuracy for each augmentation method on the test set. Each category represents a model trained with a different augmentation method: Baseline = raw data; S+S = stretching and shearing; Rotate = random rotation; Noise = additive Gaussian noise; Jitter = random adjustment of brightness/contrast; Flip = random horizontal and vertical flipping; Blur = random Gaussian blur; AI-Gen = additional training images generated using generative AI. Asterisks denote statistical significance relative to the Baseline model based on Mann-Whitney U test: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***), ns = not significant. Exact average test accuracies (%): Baseline: 80.15, S+S: 81.67, Rotate: 82.85, Noise: 66.55, Jitter: 80.49, Flip: 81.84, Blur: 77.91, AI-Gen: 50.59.

Figure 4 presents the average test accuracies of models trained with different augmentation methods, along with their statistical significance relative to the baseline model, as determined by Mann–Whitney U tests. Error bars were omitted as the variability was too small to be visually distinguishable. For the baseline model, 20 independent training runs were conducted. For the other augmentation methods, the number of runs varied between 6 and 21.

Among the augmentation methods, Rotate, S+S, and Flip showed statistically significant improvements over the baseline model. In contrast, models trained with Gaussian noise and AI-generated images showed significantly worse performance. No significant difference was observed for Jitter and Blur, despite using distributions from 15 models for each augmentation method.

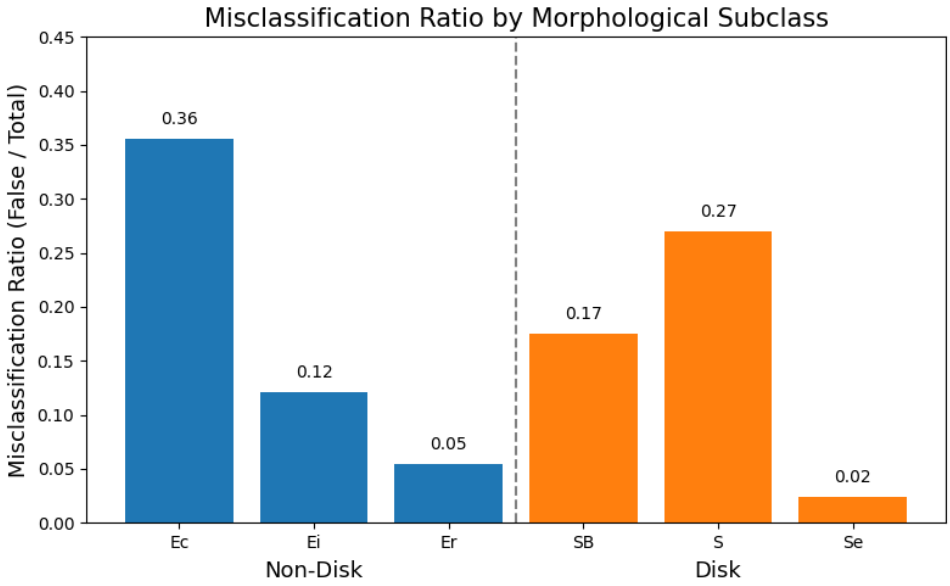

Figure 5: Misclassifications ratios were computed for each subclass based on the model's binary classification of disk and non-disk categories. E denotes non-disk galaxies, with subclasses c, i, and r representing elongated (cigar-shaped), intermediate and completely round shape respectively, ordered by increasing roundness. SB indicates barred disk galaxies and Se refers to edge-on disk galaxies. S denotes disk galaxies that are neither barred nor viewed edge-on.

These results indicate that only augmentation methods which directly altered the shape of the galaxy were effective. In contrast, augmentations that did not modify shape — such as those altering brightness or texture — did not lead to improvements in classification accuracy. This further suggests that colour information may be of limited value for morphological classification tasks. An exception to this trend was the use of AI-generated images, which performed significantly worse, likely due to their limited diversity rather than the nature of the augmentation itself. The limitations of AI-generated images in this experiment are discussed further in Section 5.

Given that Rotate, S+S, and Flip produced statistically significant improvements, these three augmentation methods were combined to train the final model. The combined augmentation methods were used in 12 independent training runs, and the model achieving the lowest validation loss was selected as the final model for detailed analysis.

The final model achieved an overall accuracy of 83.28%, with 88.07% accuracy for classifying

non-disk galaxies and 79.19% for disk galaxies. During training, the model displayed signs of underfitting, regardless of modifications to its capacity. This suggests that the ceiling effect on accuracy was due to the training data itself, rather than a limitation of the model. This issue is discussed further in Section 5.

Figure 5 shows misclassification ratios by the morphological subclass. Within the non-disk class, Ec exhibited the highest error rate, while Er had the lowest, indicating a clear trend of decreasing misclassification as the shape became rounder. The opposite pattern was observed in the disk class, where S showed the highest error rate and Se the lowest.

Although the model performed binary classification and did not explicitly distinguish between morphological subclasses, the misclassification ratios revealed underlying trends. In ambiguous cases, the model appeared to default to a single category that statistically offered the best chance of minimising loss. For example, galaxies with elongated appearances were often misclassified into the disk class, suggesting that the model sought to secure partial credit by consistently assigning such galaxies to a single category, possibly corresponding to Se, rather than risk a completely incorrect prediction. A similar tendency was observed for rounder galaxies, which were frequently classified as non-disks, potentially aligning with Er. While these subclass associations are not definitive, the results suggest that the model relied primarily on apparent 2D shape, rather than recognising the underlying 3D structure of the galaxy.

5 Discussion

5.1 Limitations of AI-Based Image Generation and Possible Improvements

Unlike other augmentation methods used in this experiment, the inclusion of AI-generated images resulted in notably poor performance. The classification accuracy for the model trained with these synthetic images dropped to 50.59%, consistent with random guessing of binary classification. This aligned with visual inspection of the generated samples. Figure 6 shows examples of AI-generated galaxy images. While individual AI-generated images appeared visually plausible, they exhibit very limited diversity. This lack of variation was likely the primary reason for the model's poor generalisation when trained on these images, and two key factors appeared to contribute to this limitation.

First, current generative diffusion models such as Stable Diffusion has inherent constraints in reproducing the full diversity of the original training dataset [8]. Even when trained on a varied dataset, these models operate within a fixed parameter space and tend to produce outputs clustered around an average representation, failing to capture spectrum of the training dataset. This limitation becomes especially pronounced when diffusion models are fine-tuned using LoRA techniques [9]. However, these constraints could potentially be mitigated through the use of

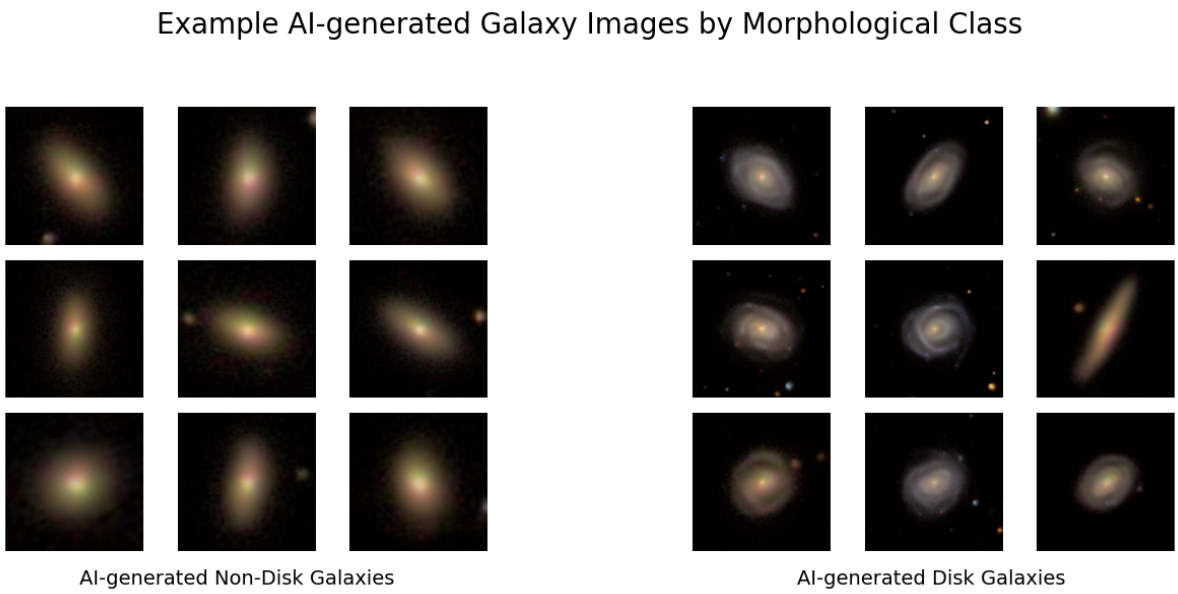

Figure 6: Examples of AI-generated galaxy images (256×256 pixels) from the two morphological classes used in this experiment. The left panel shows AI-generated non-disk galaxies, while the right panel shows AI-generated disk galaxies.

FouRA (Fourier Low Rank Adaption) as suggested in recent work [9]. Additionally, training a conditional diffusion model from scratch, rather than fine-tuning an existing one, may yield more diverse outputs [10], although this approach would require significantly more computational resources than using LoRA or its variants.

Second, the use of simple, single-word captions ("disk galaxy" or "non-disk galaxy") during LoRA training likely exacerbated the problem. In typical generative workflows, caption diversity helps the model to produce a broader set of outputs. The Galaxy Zoo 2 dataset includes detailed morphological annotations, which could have been used to create more descriptive captions. This might have encouraged the model to learn a wider range of the morphological features.

Although the AI-generated images failed to improve the performance in this experiment, the approach still holds potential with usage of FouRA and improvements in prompt design. Generative augmentation method may become a viable technique for improving generalisation.

5.2 Limitations Due to Data Ambiguity

While a maximum test accuracy of 83.28% achieved using augmentation methods that showed significant improvement, the model consistently displayed signs of underfitting during training. Regardless of architectural modification, such as increasing depth or removing regularisation methods like dropout and weight decay, the training accuracy did not exceed around 84%. This indicates that the model capacity was not the limiting factor. Instead, the constraint appeared

to stem from quality and ambiguity of the training data itself.

An analysis of original vote rates of Galaxy Zoo 2 project supports this view. The mean vote rates distinguishing between disk and non-disk galaxies across all images was $(81.0 \pm 0.2)\%$, whereas this dropped to $(66.9 \pm 1.5)\%$ for images the model misclassified. This difference indicates that the model struggled particularly with images that humans also found ambiguous.

This interpretation agrees with findings from previous studies as well. For example, Schneider et al. (2023), reported a similar ceiling accuracy of around 84% when using unfiltered Galaxy Zoo 2 classifications for comparable tasks, even with pretrained models [11]. By contrast, significantly higher accuracies, exceeding 99%, have been achieved with Galaxy Zoo 1 classification, but only when training on filtered data with the vote rates exceeding 80%, as reported by Cheng et al (2020). However, even within these filtered samples, the study reported that approximately 2.5% of galaxies with vote rates above 80% were still likely to be mislabelled. They attributed these misclassifications to the limited resolution of SDSS imaging, and identified them by comparing to higher-resolution images from Dark Energy Survey (DES) [12]. Since the Galaxy Zoo 2 project was also based on SDSS data, it is likely that substantial number of images in the dataset were similarly mislabelled.

Taken together, these results strongly suggest that the observed performance ceiling was not a flaw in the model architecture, but a reflection of the inherent limitations in the datasets. Therefore, data curation should be prioritised to increase the classification accuracy. In addition, in future developments of automated galaxy morphology classifiers, it would be advantageous to restrict training to high-confidence labels and to treat ambiguous cases separately through human review.

6 Conclusion

In this experiment, to evaluate the effectiveness of various augmentation methods, each method was applied independently on the training data. Among the augmentation methods, random rotation, stretching and shearing, and flipping showed statistically significant improvements over the baseline model. In contrast, models trained with Gaussian noise and AI-generated images showed significantly worse performance. No significant differences were observed for random adjustment of brightness/contrast, and Gaussian blur. These results indicated that only augmentation methods which directly altered the shape of the galaxy were effective.

Given that random rotation, stretching and shearing, and flipping produced statistically significant improvements, these three augmentation methods were combined to train the final model. The final model achieved an overall accuracy of 83.28%, with 88.07% accuracy for classifying non-disk galaxies and 79.19% for disk galaxies. Although a maximum test accuracy of 83.28%

was achieved using the augmentation methods that showed significant improvement, the model consistently displayed signs of underfitting during training. Regardless of architectural modification, such as increasing depth or removing regularisation methods like dropout and weight decay, the training accuracy did not exceed around 84%. This indicated that the model capacity was not the limiting factor. Instead, the constraint appeared to stem from quality and ambiguity of the training data itself. An analysis of original vote rates of Galaxy Zoo 2 project and the previous studies supported this view. Therefore, data curation should be prioritised to increase the classification accuracy. In future developments of automated galaxy morphology classifiers, it would be advantageous to restrict training to high-confidence labels and to treat ambiguous cases separately through human review.

In addition, unlike other augmentation methods used in this experiment, the inclusion of AI-generated images resulted in notably poor performance because of its very limited diversity. Although the AI-generated images failed to improve performance in this experiment, the approach still shows potential with the use of Fourier Low Rank Adaptation (FouRA) and improvements in prompt design.

References

[1] E. Hubble, *The Realm of the Nebulae*. New Haven, CT: Yale University Press, 1936.

[2] T. D. Oswalt and W. C. Keel, eds., *Planets, Stars and Stellar Systems: Volume 6: Extragalactic Astronomy and Cosmology*, vol. 6 of *Astrophysics and Space Science Library*. Dordrecht: Springer, 2013.

[3] C. J. Conselice, "The evolution of galaxy structure over cosmic time," *Annual Review of Astronomy and Astrophysics*, vol. 52, pp. 291–337, 2014.

[4] K. W. Willett, C. J. Lintott, S. P. Bamford, K. L. Masters, *et al.*, "Galaxy Zoo 2: detailed morphological classifications for 304,122 galaxies from the Sloan Digital Sky Survey," *Monthly Notices of the Royal Astronomical Society*, vol. 435, no. 4, pp. 2835–2860, 2013.

[5] Galaxy Zoo Collaboration, "Galaxy Zoo data release." `https://data.galaxyzoo.org/`, 2023. Accessed: 2025-06-22.

[6] Bmaltais, "kohya-ss: Training scripts for Stable Diffusion models." `https://github.com/bmaltais/kohya_ss`, 2023. Accessed: 2025-06-22.

[7] S. AI, "Stable Diffusion v1.5." `https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5`, 2022. Accessed: 2025-06-22.

[8] M. Dombrowski, W. Zhang, S. Cechnicka, H. Reynaud, and B. Kainz, "Image generation diversity issues and how to tame them," *arXiv preprint arXiv:2411.16171*, 2024.

[9] S. Borse, S. Kadambi, N. P. Pandey, K. Bhardwaj, V. Ganapathy, S. Priyadarshi, R. Garrepalli, R. Esteves, M. Hayat, and F. Porikli, "Foura: Fourier low-rank adaptation," in *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS)*, 2024. https://proceedings.neurips.cc/paper_files/paper/2024/file/83960718b4d12f799985206f1b1cf00f-Paper-Conference.pdf.

[10] C. Ma, Z. Sun, T. Jing, Z. Cai, Y.-S. Ting, S. Huang, and M. Li, "Can ai dream of unseen galaxies? conditional diffusion model for galaxy morphology augmentation," *arXiv preprint arXiv:2506.16233*, 2025.

[11] J. Schneider, D. C. Stenning, and L. T. Elliott, "Efficient galaxy classification through pretraining," *Frontiers in Astronomy and Space Sciences*, vol. 10, 2023.

[12] T.-Y. e. a. Cheng, "Optimizing automatic morphological classification of galaxies with machine learning and deep learning using dark energy survey imaging," *Monthly Notices of the Royal Astronomical Society*, vol. 493, no. 3, pp. 4209–4228, 2020.