Galaxy Morphology Classification Using Vision Transformer Architecture

1st Zeerak Baig
dept. Academy of Computer Science and Software Engineering
Johannesburg, South Africa
217026768@student.uj.ac.za

***Abstract*—Galaxy morphology is a byproduct of galaxy formation, environment interaction, internal disturbances, active galactic nuclei, dark matter effects, and the star formation histories of galaxies. Morphological categorization is a crucial piece of knowledge to define samples of galaxies and analyze the universe's large-scale structure. The main difficulty is developing a trustworthy method for doing morphological estimates from galactic pictures. This article proposes two deep learning approaches for galaxy morphology classification using images of various galaxies. Our first approach explores the usage of a vision transformer, whereas the second approach will use a convolutional neural network with EfficientNet B7 as a backbone architecture. The article will comprehensively compare the two deep learning methods for a comparative study. Such systems can substantially improve galaxy classification within large data sets and contribute greatly in the domain of observable cosmology.**

***Index Terms*—Vision Transformer, Convolutional Neural Network, Morphology**

I. Introduction

The morphological classification is the most fundamental data in observational cosmology for building galaxy catalogues. Galaxies with a dominating bulge component are known as Early-Type Galaxies. In contrast, galaxies with a prominent disk component are known as Late-Type Galaxies, according to the initial classification scheme developed by Hubble in 1926 and 1936. Because of their conspicuous spiral arms, Late-Type Galaxies are sometimes referred to as spiral galaxies. In contrast, Early-Type Galaxies are frequently referred to as elliptical (E) galaxies because of their more straightforward ellipsoidal shape and lack of structural differentiation (less information). According to more precise classifications, spirals fall into two categories: barred (SB) and unbarred (S) galaxies. The strength of these two groups' spiral arms can also be used to separate them further. The morphological categories can be categorized by a number called T-Type: ETGs have T-Type 0, and LTGs have T-Type 0. T-Type considers ellipticity and the strength of the spiral arms but does not account for the existence or absence of the bar characteristic in spirals [1].

The inherent, structural, and environmental characteristics of galaxies are revealed by their optical appearance. These characteristics show the age of galaxies, the history of galaxy formation, and interactions with other galaxies. Much of our understanding of galaxy morphological classification since Hubble's ground-breaking approach rests on visual observation [2].

Identify applicable funding agency here. If none, delete this.

Machine learning, specifically deep learning, has recently made considerable strides in image classification and recognition. Over the past few years, various machine learning techniques have been applied for Galaxy morphology classification. Our study will use an image classification architecture known as the vision transformer for galaxy morphology classification. The article will also use a convolutional neural network with EfficientNet B7 architecture as the backbone for morphology classification to provide a comparative analysis.

II. Background

Galaxies come in a variety of forms, dimensions, and hues. Galaxies need to be categorized to comprehend how the morphologies of galaxies connect to the physics that creates them. Therefore, classifying galaxy shapes is essential in understanding galaxy creation and development. Edwin Hubble first proposed the "Hubble Sequence" in 1926, also known as the "Hubble Tuning Fork," utilizing visual analysis of fewer than 400 galaxy photos. He also divided galaxies into three main categories: elliptical, spiral, and irregular. We still employ the "Hubble Sequence" nowadays. Astronomers have long updated the Hubble classification system and classified galaxies using visual observation [3].

Large-scale surveys like the Sloan Digital Sky Survey (SDSS) have produced many galaxy pictures in recent years. Astronomers' attempt to categorize this large number of photos is futile and time-consuming. The main difficulty is developing a trustworthy method for doing morphological estimates from galactic pictures.

A. Existing Works

Before diving straight into our approach, looking at existing works related to galaxy morphology classification using various deep-learning methods is essential. Kalvankar et al. studied the usage of efficient nets and their applications in galaxy morphology classification. They explored the usage of efficient nets to predict the vote fractions of the 79975 testing images from the Galaxy Zoo dataset. This project used Efficient net B05 architecture to classify galaxies into seven different classes with an accuracy of 93.7 per cent and an f1-score of 88.75 per cent [4].

Zhang et al. proposed a method for galaxy morphology classification using Few-shot learning. Their study used data that was split into five categories for the Galaxy Zoo Challenge Project on Kaggle following the accompanying truth table. By categorizing the data set above using supervised deep learning based on AlexNet, VGG 16, and ResNet 50 trained with various volumes of training sets individually, as well as few-shot learning based on Siamese Networks. Their findings showed that few-shot learning often yields the best accuracy, with the greatest gain over AlexNet being 21 per cent when the training sets comprise 1000 pictures. In addition, few-shot learning takes about 6300 photos for training, whereas ResNet 50 needs about 13,000 images to guarantee that accuracy is at least 90 per cent [5].

Lin et al. used an Efficient vision transformer for galaxy morphology classification. For the first time, their work investigated the application of a Vision Transformer (Vit) for categorizing galaxy morphology. They demonstrated that ViT could get competitive results with CNNs and is particularly effective at classifying smaller and fainter galaxies. Their method was able to achieve an overall accuracy of 80.55 per cent [2].

III. Experiment Setup

This study aims to efficiently group different galaxy morphologies into the appropriate classifications. For Galaxy morphology classification, our first method employs a transformer-based architecture. The Vision Transformer, often known as ViT, is a method of classifying pictures that gives some areas of the image a Transformer-like shape. A typical Transformer encoder receives the sequence of vectors that results from dividing a picture into fixed-size patches, the linear embedding of each patch, and the addition of position embeddings. The conventional method of adding an extra, teachable "classification token" to the sequence is utilized to do classification. Our second strategy is an adaptation of our first pipeline, depicted in Figure 2, in which a EfficientNet B7 architecture is utilized as a classifier rather than the vision transformer for side-by-side analysis.

A. Dataset

The study made use of a single dataset, known as the Galaxy10 DECals dataset. This dataset is considered to be an improved version of the original Galaxy 10 dataset. Approximately 270k SDSS galaxy pictures were classified by volunteers for the original Galaxy10 dataset using Galaxy Zoo (GZ) Data Release 2, and 22k of those images were chosen for inclusion in 10 major classes based on volunteer votes.Later, GZ used photos with significantly higher resolution and image quality from DESI Legacy Imaging Surveys (DECals). Galaxy10 DECals integrated all three (GZ DR2 with DECals pictures instead of SDSS photos and DECals campaign ab, c) to produce around 441,000 distinct galaxies covered by DECals, of which 18,000 were chosen in 10 broad classes by volunteer votes with more stringent filtering. Edge-on Disk with Boxy Bulge class, which had just 17 pictures in the original Galaxy10, was dropped, and the 10 broad classes of the Galaxy10 DECals were slightly modified to make each class more different from the others. The dataset has 17736 256x256-pixel colored galaxy pictures in the g, r, and z bands that are divided into 10 classes [6].

B. Evaluation Metrics

The study will present several measures to assess the classifiers' precision. When the classes are very uneven, precision and recall are valuable indicators of prediction success. Recall quantifies how many relevant results are retrieved, while precision values come from relevance in information retrieval [12]. The precision-recall curve represents the tradeoff between accuracy and recall for different thresholds. A substantial area under the curve implies strong memory and precision, whereas high accuracy recommends a low false-positive rate, and high recall suggests a low false-negative rate. High scores indicate that the classifier produces accurate findings and that most positive results are positive.

IV. Method

The three essential components of the approach are preprocessing, training, and validation, as seen in Figure 1. We use the photos in the training set to train the vision transformer model after selecting the dataset. The picture data is standardized and the sizes are modified for homogeneity during the preparation step. Without utilizing any pre-trained models, the conventional vision transformer that was developed in this work was put to the test against a Efficient Net B7 CNN architecture to classify galaxy morphology.

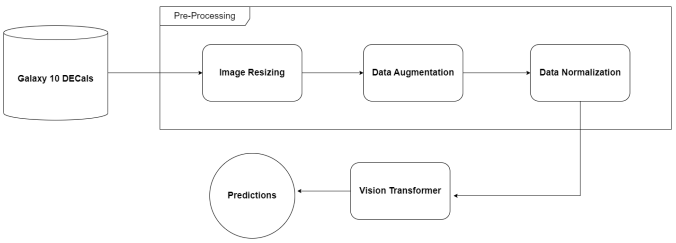

Fig. 1. Proposed Method.

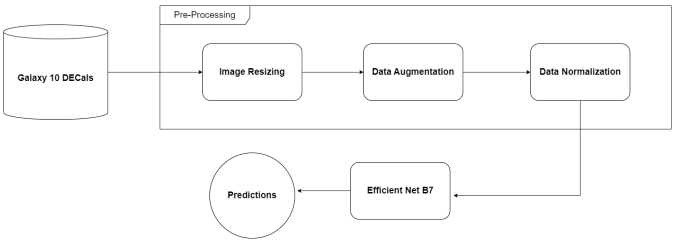

Fig. 2. Pipeline variation with Efficient Net B7 Architecture.

A. Data Preprocessing

The dataset used had a total of 21785 galaxy images. The distribution of galaxy images among various classes was quite imbalanced, as shown in Figure 3 below. Disk-face and Non-spiral, smooth, completely round, and smooth in-between round galaxies are overrepresented compared to the rest of the classes. The training set used 19606 images, and the test set for our model used 2179 images belonging to 10 different classes. During the data preprocessing stage, galaxy images go through normalization, resizing by 72x72 pixels, and data augmentation before being saved in a dataset that will be used by both the vision transformer and the Efficient Net B7.

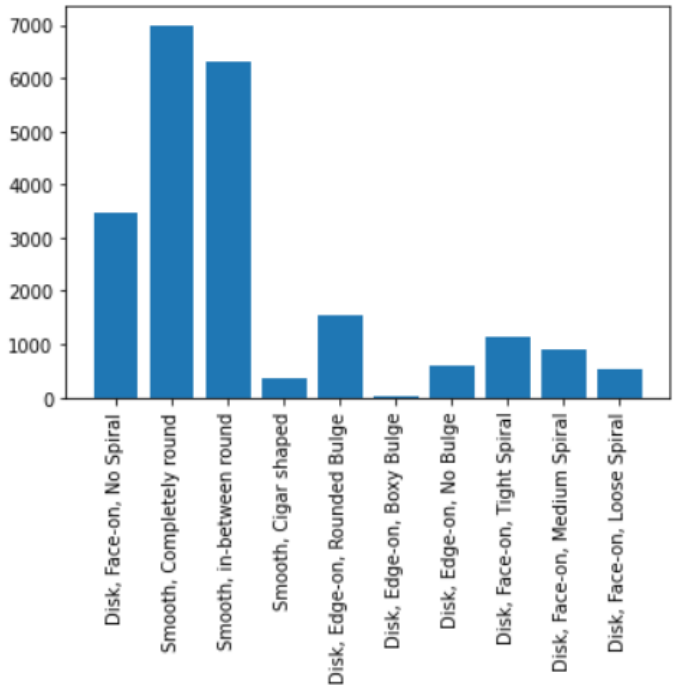

Fig. 3. Image count per class of galaxy morphology.

B. Vision Transformer Architecture

Vision Transformer is an architecture that is built on the original Transformer. The original Transformer is a prevalent architecture because it performs well in NLP applications like machine translation. Without a recurrent network, the Transformer's architecture of encoders and decoders allows it to process sequential input in parallel. The performance of Transformer models has been dramatically influenced by the self-attention mechanism, which is hypothesized to capture long-range connections between the sequence's parts.

The suggested Vision Transformer is an effort to expand the application of the conventional Transformer to picture categorization. With no integration of a data-specific architecture, the key objective is to generalize them to other modalities outside the text. The encoder module of the Transformer is used explicitly by Vision Transformer to carry out classification by mapping a series of picture patches to the semantic label. The attention mechanism used by the Vision Trans-former allows it to attend across various parts of the picture and integrate information throughout the whole image, in contrast to standard CNN designs that often use filters with a small receptive field [8].

Figure 4 displays the whole end-to-end architecture of the model. A final head classifier, an encoder, and an embedding layer generally make up this system. In the first phase, non-overlapping patches are created from a picture X from the training set (we omit the image index I for simplicity). The Transformer sees each patch as a distinct token. So, with an image of size X, where c is the number of channels, h is the height, and w is the width, we extract patches from each dimension c x p x p. This creates a series of patches with lengths (x1, x2,..., xn) where n=hw/p2. A 16 by 16 or 32 by 32 patch size is typically selected, with a lower patch size resulting in a longer sequence and vice versa [9].

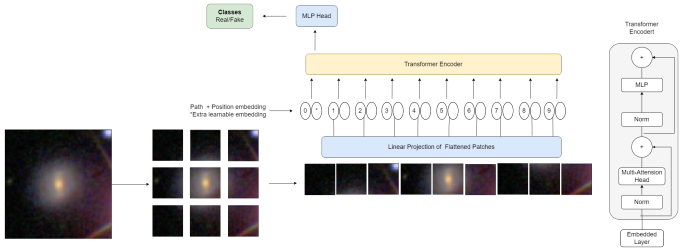

Fig. 4. Vision Transformer Architecture.

C. Efficient Net B7 Architecture

Before feeding our convolutional neural network with train and test samples, image samples must be preprocessed. The images are firstly resized to 69 by 69 pixels, and we convert them from greyscale to RGB space by repeating the intensity values across all three channels. The process then reads the image in RGB format and applies pixel normalization.

Once the image pre-processing has been completed, our convolutional neural network is ready to accept the input data. Before feeding data to the CNN, the training data goes through data augmentation stage, which increases the diversity of dataset without the need to collect more data.

The sequential model developed by Keras is used in the suggested investigation. In our model, Efficient Net B07 is the top layer. EfficientNet is a convolutional neural network design and scaling approach that scales all depth, breadth, and resolution parameters uniformly using a compound coefficient. In contrast to conventional practice, which scales these elements freely, the EfficientNet scaling technique equally increases network breadth, depth, and resolution with a set of predetermined scaling coefficients [15]. A two-dimensional Global Average Pooling layer follows our architecture. The goal of the global average pooling layer is to replace the fully linked layers in a typical CNN [10].Instead of building completely connected layers on top of the feature maps, we average each feature map and send the resulting vector directly

to a softmax layer. A dropout layer with a 20 per cent dropout rate is then added. The technique ignores or drops out a certain number of neurons in the network at random. Finally, for a binary classification job, we have a fully linked layer with a sigmoid activation function. Figure 5 below shows an efficient Net B07 design.

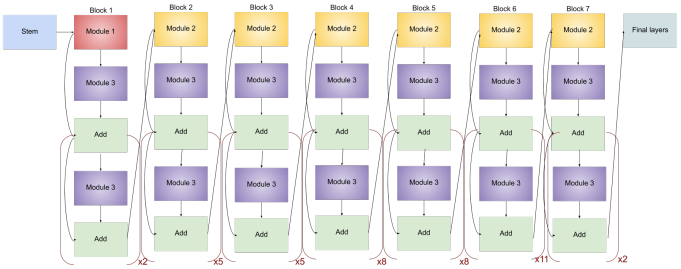

Fig. 5. Efficient Net B7 Architecture.

V. Results

A. Vision Transformer Architecture

According to our results, the visions transformer architecture achieved an overall accuracy of 82.4 per cent. The vision transformer classifier achieved a precision of 90.4 per cent and an average recall of 88.8 percent. Figure 6 and 7 below demonstrates the training and validation curves for our classifier.

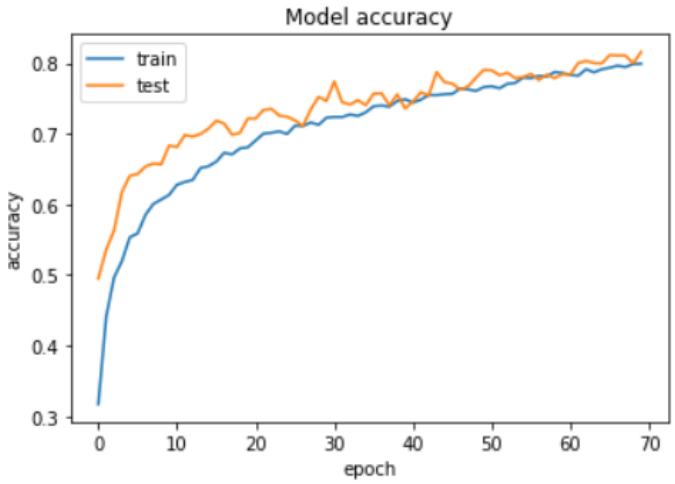

Fig. 6. Training and validation accuracy curves for Vision Transformer.

Figure 8 below is the confusion matrix for the vision transformer classifier. According to our results, the vision transformer classifier correctly classified 1786 galaxies into their respective classes, whereas 393 galaxy morphology images were incorrectly classified.

B. Efficient Net B7 Architecture

A separate architecture was used for classification reasons in our original pipeline. Using the same picture pre-processing

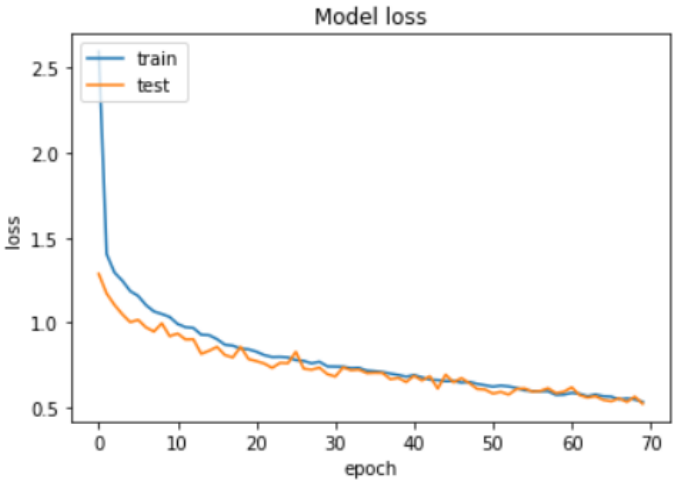

Fig. 7. Training and validation loss curves for Vision Transformer.

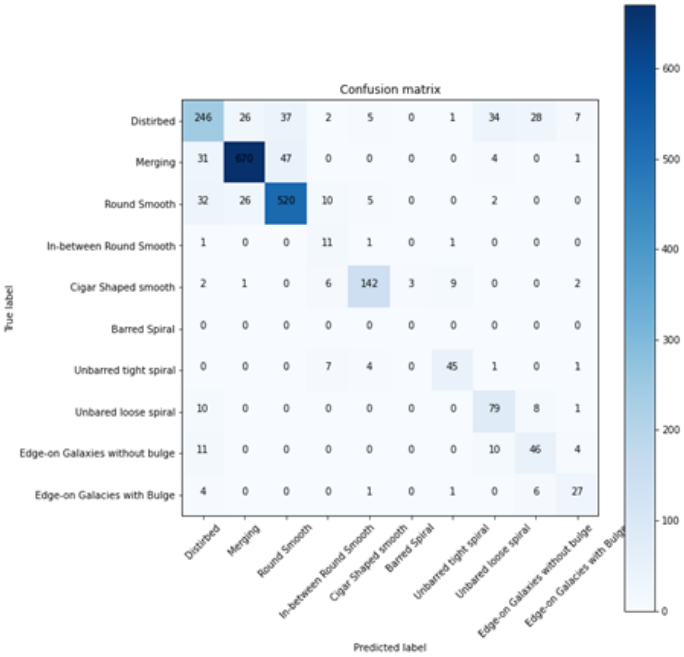

Fig. 8. Confusion matrix for vision transformer architecture.

and normalization approaches, we input the images into a convolutional neural network using an Efficient Net B07 architecture. The classifier was able to achieve an accuracy of 68.15 per cent, which is significantly lower than the accuracy of our vision transformer architecture. The vision transformer architecture outperformed the Efficient Net architecture by 16.04 per cent. Figures 9 and 10 below show the training and validation curves for the efficient net classifier.

Figure 11 below is the confusion matrix for the Efficient Net B7 classifier. According to our results, the classifier correctly classified 1485 galaxies into their respective classes, whereas 692 galaxy morphology images were incorrectly classified.

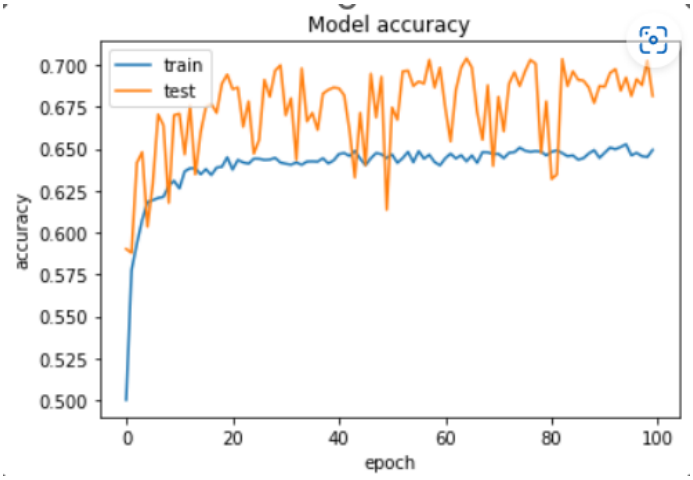

Fig. 9. Training and validation and validation accuracy of Efficient Net B7 architecture.

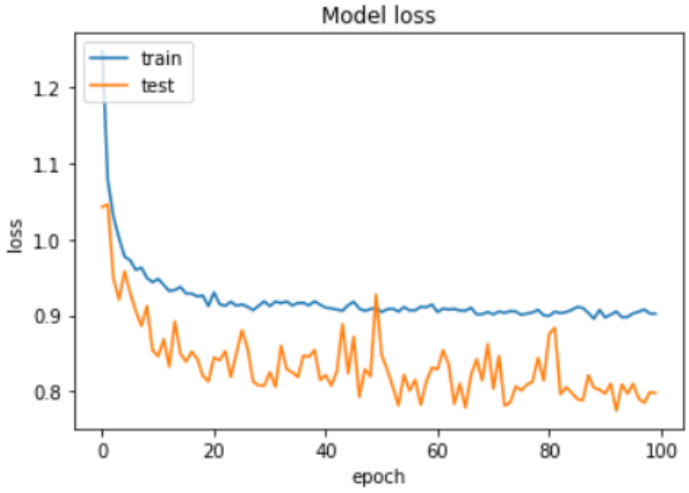

Fig. 10. Training and validation loss curves for Efficient Net B7 architecture.

VI. Discussion

The results indicate that our initial pipeline with a vision transformer as a classifier outperformed the Efficient Net B7 classifier by nearly 16.04 per cent in accuracy. Even though our vision transformer implementation achieved a decent accuracy of 82.4 per cent, it could not outperform any of the previous research outputs mentioned in section 2.

Comparing our highest accuracy with some of the earlier works, we can see that the results of kalvankar et al. outperformed our model by approximately 11.3 per cent. Zang et al. revealed that few-shot learning frequently produces the best accuracy, with the most significant improvement over AlexNet being 21 per cent when the training sets contain 1000 images. In addition, while ResNet 50 requires roughly 13,000 images to ensure accuracy of at least 90 per cent, few-shot learning only needs about 6300 shots for training. Their research also beat our vision transformer implementation by 7.6 per cent. Finally, we can compare our results to Lin et al.'s efficient

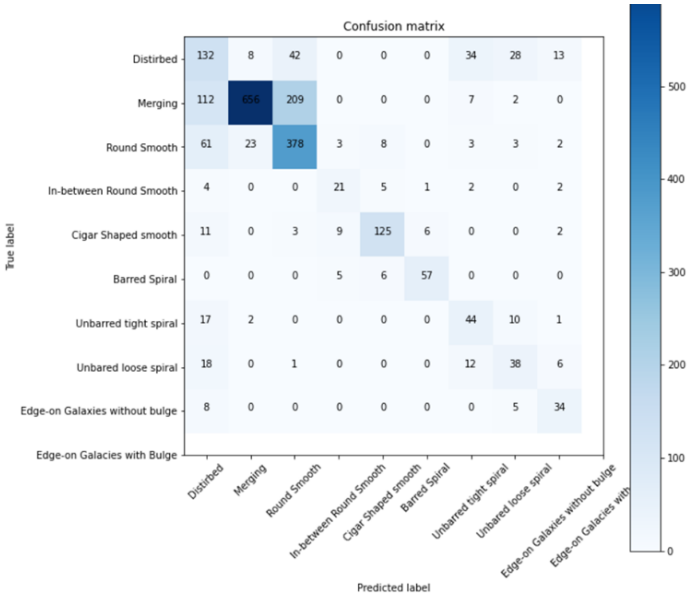

Fig. 11. Confusion matrix for Efficient Net B7 Architecture.

vision transformer implementation, which achieved an overall accuracy of 80.55 per cent for classifying fainter galaxies. We can say that our model outperformed the efficient vision transformer model by 1.9 per cent.

VII. Conclusion

Two different classifiers were presented for galaxy morphology classification for a comparative study. Our initial approach made use of a vision transformer as the classifier. In contrast, a variation of the same pipeline used an Efficient net B07 architecture for galaxy morphology classification into ten different classes. The study presented two deep learning approaches with various advantages and a decent accuracy score. The vision transformer model achieved an accuracy of 82.4 per cent, whereas the CNN-based architecture acquired an accuracy of 68.15 per cent. Understanding the origin and development of galaxies and their sub-components as a consequence of brightness, environment, star formation, and galaxy assembly throughout cosmic time requires a study of the morphology of galaxies. It takes a lot of galaxy samples and automated morphology measurement techniques to separate the factors that influence galaxy evolution and shape [11]. Therefore we need tools that can efficiently classify galaxy morphology with high accuracy. The results obtained in this study can be considered adequate. However, we can achieve better results with more data and hyperparameter tuning.

References

[1] P. H. Barchi, R. R. de Carvalho, R. R. Rosa, R. A. Sautter, M. Soares-Santos, B. A. D. Marques, E. Clua, T. S. Gonçalves, C. de Sá-Freitas, and T. C. Moura, "Machine and deep learning applied to Galaxy Morphology - A Comparative Study," Astronomy and Computing, vol. 30, p. 100334, 2020.

[2] J. Y.-Y. Lin, S.-M. Liao, H.-J. Huang, W.-T. Kuo, and O. Hsuan-Min Ou, Galaxy Morphological Classification with Efficient Vision Transformer, Oct. 2021.

[3] X.-P. Zhu, J.-M. Dai, C.-J. Bian, Y. Chen, S. Chen, and C. Hu, “Galaxy morphology classification with deep convolutional Neural Networks,” Astrophysics and Space Science, vol. 364, no. 4, 2019.
[4] S. Kalvankar, H. Pandit, and P. Parwate, Galaxy Morphology Classification using EfficientNet Architectures, vol. 1, Aug. 2020.
[5] Z. Zhang, Z. Zou, Y. Chen, and N. Li, Classifying Galaxy Morphologies with Few-Shot Learning, vol. v2, Feb. 2022.
[6] H. Leung and J. Bovy, “Galaxy10 decals dataset,” Galaxy10 DECals Dataset - astroNN 1.1.dev0 documentation, 21-Oct-2022. [Online]. Available: https://astronn.readthedocs.io/en/latest/galaxy10.html. [Accessed: 22-Oct-2022].
[7] J. Davis and M. Goadrich, “The relationship between precision-recall and ROC curves,” Proceedings of the 23rd international conference on Machine learning - ICML ’06, 2006.
[8] Y. Bazi, L. Bashmal, M. M. Rahhal, R. A. Dayil, and N. A. Ajlan, “Vision Transformers for Remote Sensing Image Classification,” Remote Sensing, vol. 13, no. 3, p. 516, 2021.
[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, vol. v2, Oct. 2020.
[10] B. Koonce, “EfficientNet,” Convolutional Neural Networks with Swift for Tensorflow, pp. 109–123, 2021.
[11] Y. Wadadekar, “Morphology of galaxies,” Galactic Astronomy, pp. 145–257, 2021.