National College of Ireland

A systematic evaluation of vision transformers for galaxy classification

MSc Research Project
Data Analytics

Pinaki Pani
Student ID: 23112573

School of Computing
National College of Ireland

Supervisor: Dr Giovani Estrada

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Pinaki Pani
Student ID:	23112573
Programme:	Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Dr Giovani Estrada
Submission Due Date:	16/09/2024
Project Title:	A systematic evaluation of vision transformers for galaxy classification
Word Count:	15231
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Pinaki Pani
Date:	16th September 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A systematic evaluation of vision transformers for galaxy classification

Pinaki Pani
23112573

Abstract

This study explores the effectiveness of Vision Transformers (ViTs) in the morphological classification of galaxies. This research utilizes the Galaxy10 Decals dataset for the deep learning tasks. The research focuses on three advanced transformer-based models—ViT Base, Swin Transformer, and DeiT Transformer alongside the conventional ResNet50 model. The Galaxy10 dataset comprises of 10 galaxy classes, serves as the benchmark for evaluating model performance. The ViT Base model is fine-tuned on the Galaxy10 dataset with weights pre-trained on ImageNet. The model demonstrated a robust performance due to its ability to capture complex relationships through multiple layers of multi-head self-attention. Similarly, the Swin Transformer is known for its hierarchical design and shifting windows, and the DeiT Transformer is enhanced with data efficiency techniques and knowledge distillation. Both the models showcased significant accuracy and precision in galaxy classification tasks.Evaluation metrics were included in this research such as precision, recall, accuracy, and F1 score. The metrics ensured a comprehensive assessment of model performances. The results indicates that the ViT Base model achieved the highest accuracy; however, a baseline CNN model performed faster. This research highlights the trade-off of Vision Transformers in the domain of astronomical image classification. It offers insights into their capability for detailed morphological analysis of images. The findings suggest that ViTs could be used as a general-purpose image classification technique, showing slightly better accuracy than ResNet50. Overall, vision transformers show superior ability to model contextual information and are promising tools for image classification.
Keywords - vision transformers, convolutional neural network, galaxy, morphology, CNN

1 Introduction

Galaxies are the main building blocks of our universe, where each one is a stunning masterpiece in the cosmic world. Understanding their morphology is very crucial to understand the cosmic mechanisms. Galaxies have very diverse shapes and forms where each one of them emits different types of light and provide unique spatial information. Such spatial information provides all kinds of insights, such as regarding their formation, evolution, and the physical processes that undergo on a regular basis. For astronomers and astrophysicists, accurately classifying these cosmic forms is a significant feat as that helps to further enhance the understanding of the universe's history[1].

Different galaxies exhibit distinct morphologies and characteristics. These morphologies can be broadly categorized into three main types: elliptical, spiral, and irregular galaxies(Lintott et al.; 2008). Accurate classification of these types further helps in the study of different fields of space exploration such as galaxy formation, the role of dark matter, and the evolution of cosmic structures. Traditional methods of galaxy classification have relied heavily on visual inspection and manual categorization. The old methods were not only labor-intensive but also very prone to human error and bias. These methods were very foundational and were effective initially. However, currently they are increasingly insufficient in handling the vast amounts of data generated by modern astronomical surveys.

[1] `https://www.zooniverse.org/about/publications#space`

Recent advancements in image-capturing technologies and pattern recognition have paved the way for automated systems. Automated methods can classify galaxy types from astronomical images with greater efficiency and accuracy. These systems are particularly valuable for large-scale surveys and deep space exploration, where vast amounts of image data are generated. The rapid development of machine learning, particularly deep learning techniques, has opened up new opportunities for astronomical data analysis. Deep Convolutional Neural Networks (CNNs) have been widely adopted for their ability to learn hierarchical feature representations from raw image data. CNNs have proven to be quite useful in terms of image classification in almost every sector. Galaxy classification is also no exception. However, the complexity of galaxy morphology requires more advanced models that can capture intricate patterns and subtle differences in the data. Better CNNs or any other technology that can provide crucial information on vast datasets are always subject to research in this field.

Vision Transformers (ViTs) represent a significant leap forward in this context(Dosovitskiy et al.; 2021). They have been very useful in multiple scenarios of pattern recognition from image datasets. CNNs rely heavily on convolutional layers to extract local features, whereas Vision Transformers use self-attention mechanisms to capture global dependencies in the image data. This approach helps them to create complex models and relationships within the data more effectively. Among the various ViT architectures, models like Swin Transformer, Data Efficient Image Transformers (DeiT), and ViT Base have shown remarkable performance in image classification tasks within different domains(Liu et al.; 2024).

In this research, the application of these advanced Vision Transformer models for galaxy morphological classification is explored. The aim is to improve the accuracy and efficiency of automated galaxy classification systems by leveraging their superior feature extraction capabilities, . Additionally, this research also compares these models with traditional CNN architectures such as ResNet50 to evaluate their relative performance and determine the most suitable approach for this task.

The importance of this research extends beyond the immediate goal of improving classification accuracy. Enhanced galaxy classification can lead to better insights into the processes governing galaxy formation and evolution. It can also help in the discovery of new and rare galaxy types while also contributing to the broader understanding of the universe.

1.1 Research Questions

The usage of vision transformers for image classification is a relatively new concept. As such, little is known about its applicability with limited amount of data in the field of galaxy morphology classification. Or, if vision transformers outperform well-established, general image processing techniques, such as CNNs in this field. There is ample scope for research in vision transformers, but for the purposes of this research report, we will focus on:

1. How do Vision Transformers compare to state of the art techniques, such as Convolutional Neural Networks? In other words, are vision transformers able to achieve comparable accuracy using limited amount of data?

2. From a range of vision transformers, which one is the most efficient in terms of accuracy and learning time for image classification?

A number of vision transformers exist, but the current work will narrow down the list to the following: Swin, Vision Transformer Base (ViT), and Data Efficient Image Transformer (DeiT)

1.2 Research Objective & Outline

In order to address the research questions, the following The following are the primary objectives of this research:

1. Data preparation – Importing the data to pre-process it followed by required data augmentations.

2. Implementing the models as per the Galaxy10 Decals dataset. The models used are Swin Transformer, ViT Base, DeiT and ResNet50.

3. Evaluate the models and check the metrics Precision, Recall, F1 Score and Accuracy.

The report is organised as follows. In Section 1 a brief introduction to the field is presented . Section 2 outlines the works that has been done within the field. Section 3 discusses how the models were implemented and trained. The Section 5 then describes the computational resources that were used to complete the research. Section 6 moves on to check the evaluation that produces the results as per the experiments done in the research. Finally, all results are discussed and concluded with the future work aspect in Section 7.

1.3 Research Gap

Vision transformers are a new way of performing image classification. There is however little information about what type of transformers are suitable for general-purpose classification tasks. In this report, a challenging dataset, with 10 classes is used to establish whether vision transformers are any better than traditional convolutional neural networks.

In other words, vision transformers have shown tremendous performance in various image classification tasks across different domains, however their comparisons do not seem fair. The training of vision transformers is done with larger datasets than traditional convolutional neural networks. Or, how does training time compare to conventional, yet amply successful CNNs?

There has been very few research that has focused on this as mostly Convolutional Neural Networks are explored in previous research for this task. This research focuses to address this issue by making use of multiple different types of vision transformers for galaxy morphology classification. On top of this generally researches focus on classification of six to seven classes of galaxies. There are more galaxy types that have not been extensively studied and explored for classification. This research also attempts to aid in this area as the dataset used here has 10 morphological classes of the galaxy images.

2 Related Work

Space Exploration has undergone massive increase in its pace to produce more and more information about the cosmic entities present all around. Many exploration such as Large Synoptic Sky Survey promises to generate abundance of images. Surveys like Galaxy Zoo and Sloan Digital Sky Survey has already generated a large number of galaxy images. These surveys are leveraged and their generated data is largely used to facilitate multiple researches in the field of space exploration. This related work section

The current space explorations being performed by scientists and explorers, promises to bring abundance of data going forward. The data generated could be massively beneficial to researchers. The Large Synoptic Sky Survey uses a Large-aperture Synoptic Survey Telescope as described in (Abell et al.; 2009). The advanced technology used for this survey, will help to bring forth thousands of images upon its completion. Other massive surveys that is already done is The Sloan Digital Sky Survey by (York et al.; 2000). Images produced from this survey provides morphological information about the galaxies. The morphological information retrieved from the survey opens up many more opportunities for research such as age, struture, formation, history and merger information on galaxies. Traditional approaches, such as those by (de Vaucouleurs et al.; 1991), and Sandage (1975), rely entirely on visual inspection. This type of inspection was extremely prone to inconsistencies and consisted of human biases. Recent advancements

focus on automated classification techniques to handle modern astronomical survey data, which comprise millions of galaxies.

2.1 Machine learning techniques used for Galaxy Morphology Classification

There has been many innovative techniques used over time to perform galaxy morphology classification. Initially, after relying on visual techniques, researchers moved on to machine learning methods for galaxy classification. Several studies used machine learning algorithms such as Naive Bayes and Random Forest to perform hierarchical classification of galaxies. For instance, (Marin et al.; 2013) conducted a study where classification was carried out following feature extraction. They also introduced the concept of calculating geometric moments. Additionally, the authors addressed class imbalance by artificially generating galaxy images using geometric transformations. Subsequently, Support Vector Machines (SVMs) were utilized for morphological classification in research conducted by (Applebaum and Zhang; 2015), which made use of the Galaxy Zoo dataset. The more detailed morphological characteristics that were captured in images can be better processed and classified using deep learning models such as Convolutional Neural Networks (CNNs) and Vision Transformers.One noteworthy contribution in this field is by (Huertas-Company et al.; 2010), who present a Bayesian automated classification method for approximately 700,000 galaxies from the SDSS DR7 dataset. In this research the authors used the machine learning technique support vector machines (SVM) to classify galaxies into four morphological types: Ellipticals (E), Lenticulars (S0), Early-type Spirals (Sab), and Late-type Spirals (Scd). They addressed the basic issues involved in manual visual classification systems. Their research methodology makes use of a combination of parameters such as color indices (g-r, r-i), shape parameters (isoB/isoA in the i-band, deVAB i), and light concentration indices (R90/R50 in the i-band) to improve classification accuracy. It is evident that their research was very robust across many perspective. They tested their models through multiple classifications with many different training samples. Upon training it was found that there were very minor variations in the probablities of the classification outputs. This shows that the classification system is very effective for all types of galaxies represented in the training set of their research. Another noteworthy research from (Barchi et al.; 2016) where the authors do in-depth study for improving galaxy morphology classification using machine learning techniques. Their research mainly targets on classifying galaxies into elliptical (E) and spiral (S) types using a set of morphological parameters. The morphological parameters considered are concentration (CN), asymmetry metrics (A3), smoothness metrics (S3), entropy (H), and a gradient pattern analysis parameter (GA). These parameters mostly are obtained from pre-processed segmented images. Their research makes use of a dataset that consists of 48,145 instances. After the dataset was preprocessed, there were 44,760 galaxies that were labeled as spiral and 3,385 were labelled as elliptical based on Galaxy Zoo classifications. The authors implemented unsupervised learning experiments with the help of techniques such as K-means and Agglomerative Hierarchical Clustering (AHC). They conducted the experiments on a dataset of 1962 instances that was also very well balanced. The authors utilized K-means clustering with 'k-means++' and 'random' initialization methods which provided almost very identical results. AHC also successfully grouped the objects into two main clusters of the galaxy types. The study demonstrated that machine learning methods such as SVM and DT, are highly effective for galaxy morphological classification. These machine learning methods achieved high accuracy. The research also helped to identify features that were most significant. Concentration parameter (CN) was identified as the most significant feature for distinguishing between spiral and elliptical galaxies. The use of grid search and cross-validation made sure that the models were well-tuned and avoided overfitting. These researches were successful to a certain degree. As in this case the models were only classifying accurately for binary classification. Other researches involves not achieving great accuracy for multiple classes. This allowed researchers to follow up with other technologies.

2.2 Classifying Galaxy Morphology utilizing Deep Learning

Over time machine learning technology evolved to use of deep neural networks. Deep neural network architectures proved to be working really well in the field of image classification. Architectures like Convolutional Neural Networks(CNNs) were crucial for further development in terms of morphological classification of galaxies. The neural network architecture of CNNs allows them to filter information from images. CNNs have convolutional layers that contains filters throughout their neural network architectures which eventually helps to extract certain patterns at certain layers. The learnable parameters within the neural network architecture helps CNNs to get better over time with more data and training. Over the years, CNNs had revolutionized galaxy morphology classification by automating the whole process and achieving better performance than manual visual classification. Studies have demonstrated that CNNs perform really well to classify galaxies into basic morphology types such as elliptical and spiral galaxy types. In their study by (Domínguez Sánchez et al.; 2018), the authors created a morphological catalogue for for approximately 670,000 galaxies from SDSS survey. This study included galaxies that has high classifier agreement from galaxy zoo for training. This helped to generalize the model better and also allowed better feature extraction. In their study they specifically made use of binary classification. They used ReLU as their activation function and the classifier layer had sigmoid activation function. In the study, the model is trained for 50 epochs with a batch size of 30. The learning rate is initially set to 0.001 but is adjusted over time. For the binary classification the model achieved high precision and recall values. Their research implementation resulted in more than 97% accuracy, hence showing how stable CNNs were.

Surveys like SDSS have been a go to source of dataset for CNNs to be used upon. Many astronomical research has been done with the help of galaxy images. The research by (Pearson et al.; 2019) explores the application of deep learning to galaxy merger identification. The authors developed a CNN architecture that they trained on both observational data from the SDSS survey and the simulated dataset obtained from the EAGLE cosmological simulation. This study targeted to test how well CNNs were able to reproduce the visual classification from observations and also physical classifications from simulations. In their research the CNN that they trained on the SDSS dataset was able to achieve an accuracy of 91.5%. Whereas the CNN that was trained on the EAGLE simulated dataset, had a lower accuracy of around 65.2%. In this study, the neural network architecture had convolutional layers along with max pooling and dropout layers. Batch normalization was also used at every convolutional layer which allowed their model to learn better than a basic CNN architecture. The model trained on the simulated data was also tested on the SDSS images, which resulted in an accuracy of about 64.4% and a 53% accuracy was obtained over the images of simulated dataset by the model trained on SDSS data.

There are other studies that went beyond binary classification where they have tried more than two-way classifications. The study by (Cavanagh et al.; 2021) trains and tests multiple CNN architectures for galaxy morphology classification. The models used in this study are used for classifying galaxies into three classes - elliptical, lenticular and spiral. The study also classifies the galaxies into four way classification as well which included the irregular or miscellaneous type. Their approach included rigorous data augmentation techniques such as cropping, rotating, and flipping to make the training dataset even better. The study also explored binary classifications between all four morphological classes. This allowed the study to identify that ellipticals and spirals are the easiest to distinguish between, while spirals and irregulars were the most difficult. In this study the authors introduced a hierarchical classification method that combines binary CNN classifiers to perform a step-by-step classification. Their method was able to get less effective accuracy for the four way classification than the three class one. This was due to the complexities that was introduced by the irregular and miscellaneous galaxy type. Their new CNN architecture, C2 performed way better than the existing models with overall classification accuracies of 83% and 81% for the 3-way and 4-way classifications, respectively.

2.3 Critical Analysis of Transfer Learning for Galaxy Morphology Classification

There has been multiple different CNN architectures that has proven to be very useful when pretrained on large labelled datasets. Transfer learning has shown benifits in all kind of domains and applications. The CNN architectures has also developed and become better with time that allows transfer learning to be more useful. The research from (Schneider et al.; 2023) shows substantial advancements with the application of convolutional neural networks (CNNs) and transfer learning. In their study the authors demonstrated a significant improvement in overall classification accuracy by utilizing the pretrained AlexNet model. Their statistical significance was greater than the non pre-trained model. Their average peak test accuracy for their pre-trained model was about 84.2%, whereas the model that wasn't pretrained had an average accuracy of 82.4%. In terms of training speed as well pre-trained model surpassed the non pretrained model greatly. Pre-trained model converged in around 155 epochs whereas the other one converged in about 367 epochs. Another such study that leveraged transfer learning and performed quite well was by (Domínguez Sánchez et al.; 2018). Their study focuses on the performance of deep learning models trained on the SDSS data when applied to images from the Dark Energy Survey (DES). The initial model was trained on SDSS and it was suitable to DES as the images have similar redshift distribution to SDSS. To improve the model performance the pretrained model was also trained on the DES image sample of about 300-500 images. Their study improved the overall accuracy from 90% to about 95% for the DES image dataset. Their result suggest that transfer learning is a suitable strategy for applying pre-trained models to new astronomical datasets. This helped to show the less need for large labeled training sets for galaxy morphology classification. They also helped to understand the issue that fine-tuning pre-trained models on a small subset of the new data can achieve very high performance levels. The performance level can even be comparable to those trained entirely on the new data. This approach turns out to be a highly efficient approach for future large-scale surveys. Other researches in the field of space exploration has also benefited from transfer learning techniques. One such research by (Ackermann et al.; 2018) found out that these deep learning methods significantly outperform previous state-of-the-art merger detection methods based on non-parametric systems. In their research they used transfer learning as a regularizer that improved their overall classification accuracy. When they used transfer learning on their overall training set they noticed a decrease in the error rate from 0.038 ± 1 to 0.032 ± 1, which is a relative improvement of 15% in their case. They also perform comparison of their method with previous automatic visual classification methods, showing significant improvements in precision, recall, and F1 score. Some other recent papers multi source datasets to perform galaxy morphology classification. Another research by (Shaiakhmetov et al.; 2021) introduces SpinalNet, a deep neural network inspired by the human somatosensory system. SpinalNet's unique approach involves a gradual input mechanism. In this mechanism intermediate layers process both new inputs and outputs from previous layers. In their study the authors used SpinalNet on the Galaxy Zoo dataset. They performed binary, three-class and ten-class classification. The binary classification considered elliptical and spiral galaxy types, where as three-class introduced the irregular galaxy type as well. The model performed really well for binary and three-class classification with about 98.2% and about 95% respectively. However, for the 10-class classification the model was able to achieve an accuracy of about 82%.

2.4 Analysis of Vision Transformers in classifying morphologies of galaxies

In 2020 the new idea of Vision transformer came into the picture. The study by (Dosovitskiy et al.; 2021) introduced vision transformer that presents a novel approach to image classification using a Transformer architecture. They showed how a vision transformer applies the transformer architecture directly to sequences of image patches. It takes sequential input as well just as

transformers do. The image patches in a ViT also contain positional embeddings retain positional information. The research shows substantial results to conclude that ViT achieves competetive performance on image recognition tasks when pre-trained on larger datasets and then fine-tuned later for smaller data tasks. Another study by (Cao et al.; 2024) focused on using deep learning techniques to classify galaxy morphologies using large-scale astronomical datasets. In their study they used the Convolutional vision Transformer (CvT), which is an improved version of the Vision Transformer model. They showed how it helped them to improve results of ViT model using a convolutional neural network. The overall accuracy achieved by them in their study was around 98%. One more notable study by (Lin et al.; 2022) explored the use of Efficient Vision Transformers (ViTs) for this purpose. In their study they mainly focues on using the Linformer model to classify galaxies. They showed that Vision transformers are able to achieve competitive results when compared with convolutional neural networks. They used the Galaxy Zoo dataset and compared the vision transformer models with ResNet-50 baseline model. The best overall accuracy they were able to achieve was 80.55% 4, whereas the best individual class accuracy achieved in their weighted-cross entropy Linformer was over 60% in each class. Even though Linformer achiever pretty good result still ResNet-50 was able to generalize better with an overall accuracy of about 85.12%.

The critical analysis of the papers over the years and the evolving technologies does indicate that it is worth checking the vision transformers in terms of classifying the galaxy morphologies with the help of the galaxy images.

3 Research Methodology

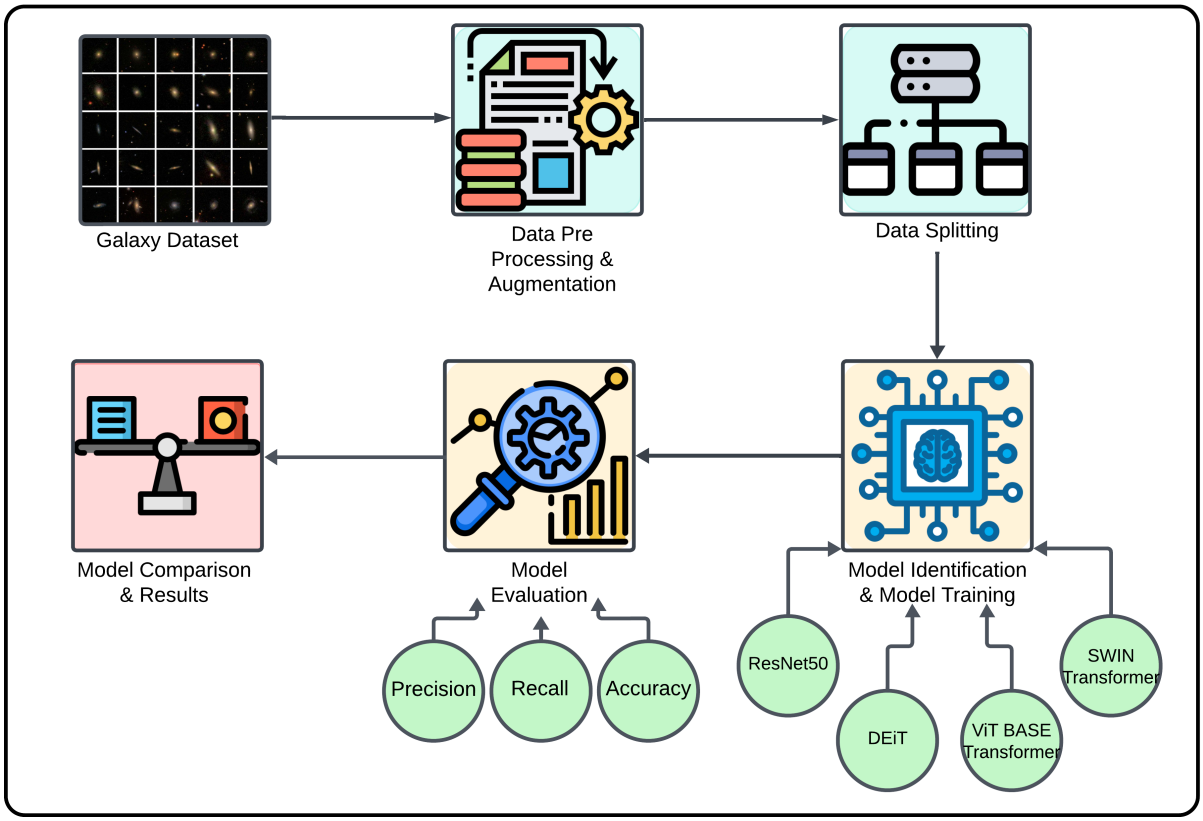

Figure 1: Research Methodology KDD Diagram

The methodology adopted in this study was structured around the Knowledge Discovery in

Databases (KDD) process. The KDD process is comprised of several critical steps, beginning with data selection, followed by preprocessing, model identification, model training, and finally model comparison along with evaluation. Each of these steps was implemented to ensure that the models trained could classify the morphological classes of galaxies in the Galaxy10 DECals dataset.

Figure 1 This research aims to identify the best Vision Transformer-based algorithm for classifying galaxy morphologies using image data as well as aims to contrast the transformer models with respect to existing convolutional neural networks. A systematic methodology as mentioned above with several key steps, including data collection, preprocessing, algorithm configuration, model training, and evaluation, is employed and the steps are described below.

3.1 Dataset Description

In the field of space exploration being able to classify galaxy images is key towards learning important information about them. Merger details, formation, evolution and movement information are few out of many crucial details that allows researchers to learn more about the galaxies and perform further research on them. The Galaxy10 DECals[2]. dataset is a curated collection of images of galaxies used primarily for research and educational purposes in the field of astronomy and astrophysics. The dataset is derived from the Dark Energy Camera Legacy Survey (DECals), which is a deep-sky astronomical survey. It is derived by the Galaxy Zoo survey, done previiously. It contains labeled images of galaxies that helps researchers to train machine learning models for automated classification of Galaxy morphology and other astronomical aspects. The dataset contains images of galaxies that are distributed among ten different morphological classes. Visual appearance and structural features of the galaxies helps to classify them into classes such as elliptical, spiral, irregular, and other unique shapes. The class distribution of the dataset and the instance count for different classes are shown in the Figure 2. Each image in

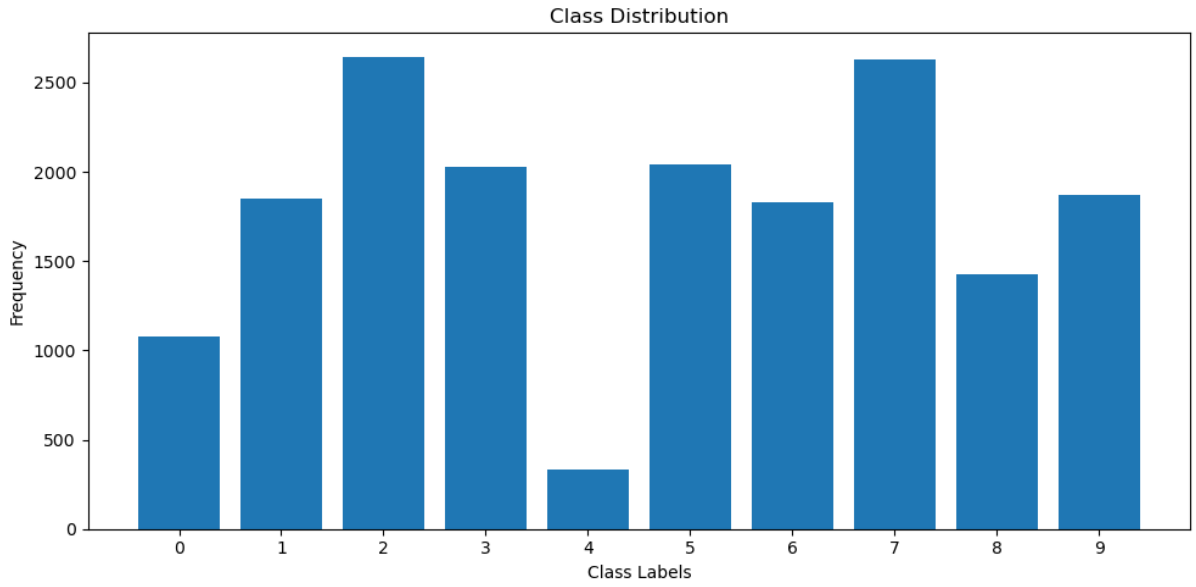

Figure 2: Class Distribution Histogram

the dataset is a square cutout centered on a galaxy. The images are typically in color with RGB channels that can help train the models and the classification process.

[2]https://astronn.readthedocs.io/en/latest/galaxy10.html#introduction

3.2 Pre-Processing of the Data

After obtaining the Galaxy10 DECals dataset from a credible source, the next crucial step was preprocessing the data. This is a very crucial step to prepre the data when one is performing deep-learning based classification. The initial raw images often contain noises, varying dimensions or some other inconsistencies. With the help of this pre-processing it can be ensured that the model learns the required features without the irrelevant inconsistencies and variances. In this research the main consideration of one such inconsistency was to consider the aspect of unnecessary information for the target classification. Some of the galaxy images contain a lot of dark background which does not provide any necessary information for the classification of galaxies. However, in certain domains the arrangement of these dark areas can be significant. For example, the pattern of black regions in a galaxy image might outline the shape or structure of celestial bodies, serving as an informative feature. Thus, while dark pixels themselves might not carry specific data, their spatial configuration relative to other pixels can be crucial for identifying and classifying objects within the image. To handle issues like that the first the images were cropped and resized to a standard dimension of 224x224 pixels. The models that were chosen were also taking a standard input of 224x224 image size so this chosen image dimension was perfect. Following this other augmentations were incorporated in the pre-processing. Random Rotation was included where the images were randomly rotated to 90 degrees to simulate multiple orientations of galaxies. The images were also flipped horizontally and vertically for mirrored version of the images. Next the main augmentation where the images were resized and cropped randomly so that the dark background that doesn't provide any information can be taken out of account while training.

3.3 Model Training Phase

Once the pre-processing is complete the Galaxy10 DECals dataset was divided into training, validation and test dataset. The distribution was done in the ration of 70:10:20 for the training, validation and test subsets respectively. The split helps to make sure that the model gets robust data for training, validation and testing. The training subset consists of images that contains data all across ten different galaxy morphology classes with respective labels. The total number of images in training subset after the split was , while the validation subset had , and the test subset had around images respectively. The Deep Learning models including vision transformers and CNN based models that were used in this research were Base Vision Transformer (ViT), Swin transformer, Data-efficient Image Transformer (DeiT) and ResNet50. All the mentioned models have been trained over the training dataset, while being monitored with validation subset while training and then evealuated using the test subset of the dataset.

3.4 Model Evaluation Phase

As soon as the models are done training, the test dataset is used to evaluate the models on different evaluation metrics. This research handles multi-class classification, so appropriate evaluation metrics are chosen for this type of tasks. Metrics such as accuracy, weighted average, precision, recall and f1 score are considered to be analysed. Precision is particularly important in classifying galaxy types because misclassifying a galaxy can lead to incorrect scientific conclusions and affect subsequent research. In astronomy, where datasets can be large, but the classes may be imbalanced (e.g., certain galaxy types being rarer than others), high precision ensures that when a model predicts a particular galaxy type, it is highly likely to be correct. This reduces the incidence of false positives, which is critical when each classification might contribute to broader scientific knowledge, such as understanding galaxy formation and evolution. This research also faced an issue with class imbalance for some classes of galaxies which is why precision turned out to be a useful metric to look into.

4 Design Specification

The design specification in this research presents the methodologies and technologies used to achieve certain accuracy for classification of galaxy morphologies. The preprocessing phase earlier involved data standardization to ensure consistency, which includes resizing images, normalizing pixel values, and augmenting data to address any existing class imbalances within the dataset. The primary aspect of this research is the implementation of Vision Transformers (ViTs). They are selected due to their ability to handle image data. The vision transformers that are used in this research are Swin Transformer, Data Efficient Image Transformer (DEiT), ViT Base Transformer and ResNet50. The network architectures and implementation are defined below.

4.1 Swin Transformer

The Swin Transformer is one of many Vision Transformers available for deep learning tasks. The Swin Transformer network architecture shows a significant advancement in the field of Computer Vision. Unlike other Convolutional Neural Netowrks (CNNs), which rely on convolutional layers to get spatial information from the images, the Swin transformer uses a heirarchical transformer structure to capture both local and global features. The attention spreads from local to global information over the image. This type of architecture consists of many stages. Each of these stages comprises of tranformer blocks in a sequence where the window size increases progressively. The increasing window size allows scalability and more efficiency within the model.
The backbone of the Swin Transformer consists of four stages, each containing several Swin Transformer layers. These layers are designed to perform self-attention within local windows. This local attention design helps to reduce the computational time and cost for the model training compared to the global self attention mechanism of a base vision transformer model. The network architecture starts with a patch partitioning layer that divides the input image into non-overlapping patches. Right after this the linear embedding layer is present that projects these non-overlapping patches into a higher-dimensional feature space. The next stages in the network makes use of the shifting window mechanism to enhance model's capability of capturing cross-window interactions.
In this research, the Swin Transformer architecture is utilized that is pre-trained on the ImageNet dataset. The ImageNet dataset includes over 1 million images across more than 1000 categories. The pre-training helps towards a robust feature extraction. The model processes images with a standard size of 224x224 pixels and then makes use of the transfer learning techniques. The Swin Transformer architecture is shown in the image 3 In addition to the Swin Transformer,

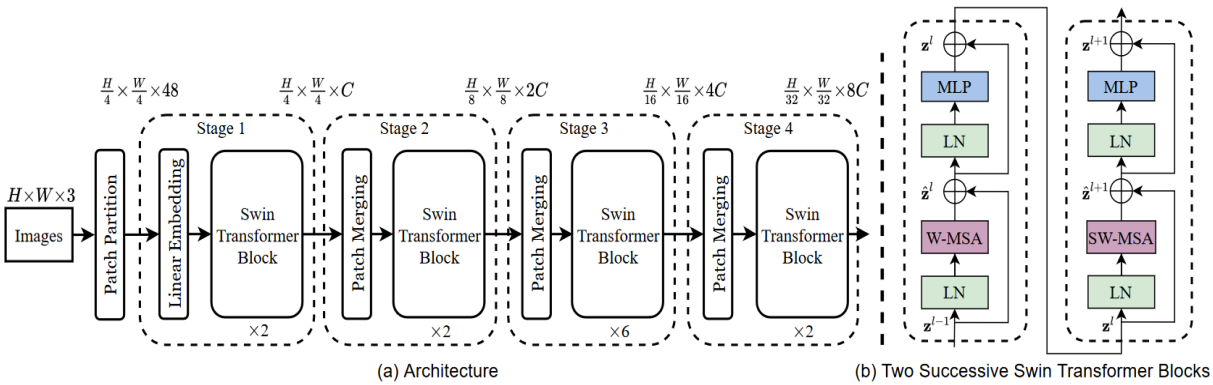

Figure 3: Swin Transformer Network Architecture, (Liu et al. (2021))

this study also employs the ResNet as baseline comparisons. They are renowned Convolutional

Neural Networks that are known for their performance in image classification tasks.

4.2 ViT Base Fine-tuned

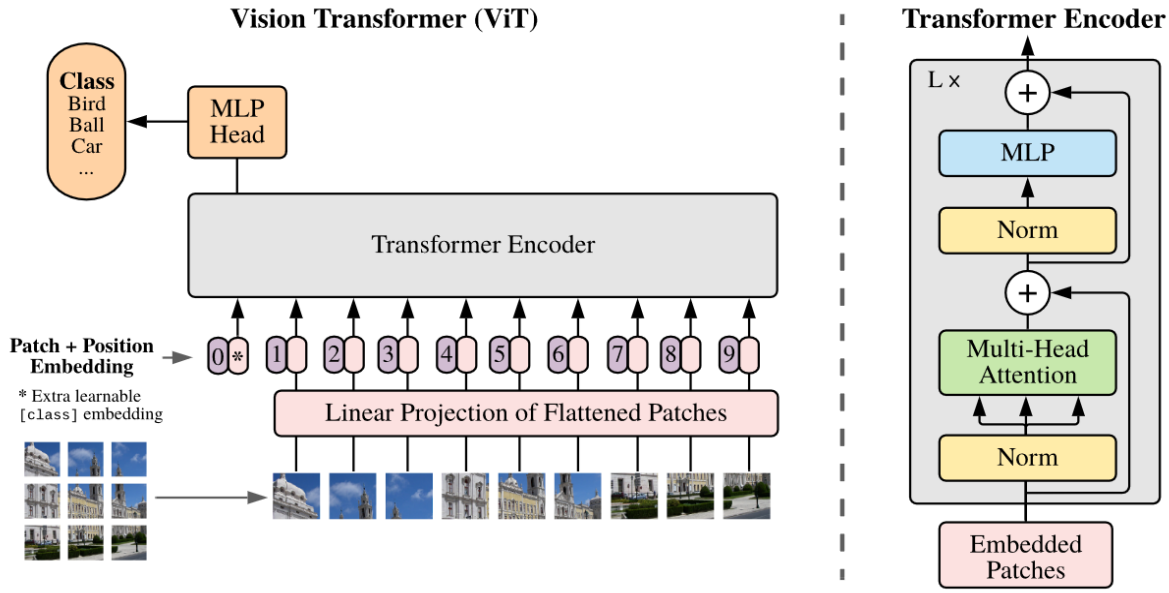

Figure 4: ViT Base Transformer Network Architecture, (Dosovitskiy et al. (2021))

The Vision Transformer (ViT) are quite new in computer vision. It moves away from traditional convolutional neural networks towards transformer-based architectures that have been highly successful in natural language processing. The ViT model employed in this research is based on the 'ViT-Base' configuration, which utilizes a patch-based approach to process images, significantly differing from traditional pixel-based convolution methods. The architecture of this model is shown in the figure 4.

In this architecture, the input image is first divided into fixed-size patches of 16x16 in this research. They are then flattened and linearly embedded into a higher-dimensional space. This embedding process transforms each patch into a vector of size 768 which is the input size of the transformer encoder. The embeddings are augmented with positional encoding to retain spatial information. This spatial information is very crucial for image-related tasks. The sequence of patch embedding forms the input to the transformer encoder. The encoder consists of multiple layers of multi-head self-attention and feed-forward neural networks. In this research the transformer encoder in the ViT model contains 12 layers. Each of these layer is equipped with 12 self-attention heads. This structure allows the model to capture complex inter-patch relationships by focusing on different parts of the image at the same time. The ViT model's strength lies in its ability to model contextual information in a global scope. This makes it particularly effective for detailed image classification tasks.

Within this research, the ViT model is fine-tuned using the Galaxy10 Decals dataset. The model is initialized with weights pre-trained on the ImageNet dataset, which comprises over a million images across 1000 categories. This pre-training enables the ViT model to leverage previously learned visual features.

Key hyperparameters used in this setup include a learning rate of $5\times1055\times105$, a step size of 5, and a gamma of 0.1 for learning rate scheduling. The model is trained for 24 epochs, with class weights applied to address any potential class imbalances. The final classification layer is a linear layer with an input feature size of 768 and an output size of 10. These ensured that the ViT model is tailored for the task of galaxy classification.

4.3 DEiT Transformer

The Data Efficient Image Transformer (DeiT) is a very efficient model in computer vision mainly for the tasks that demand high accuracy with limited data. DeiT uses a transformer-based framework to achieve extremely well performance in image classification. The DeiT architecture is built upon the foundation of the Vision Transformer (ViT). It employs some enhancements to the ViT architecture to improve data efficiency and overall model performance. The core structure of DeiT consists of a series of transformer blocks. Each of these blocks implement multi-head self-attention mechanisms and feed-forward neural networks. These blocks are very good at capturing complex patterns and relationships within image data.

First, the input image is divided into a grid of non-overlapping patches of size 16x16 pixels. Each of these patch is embedded into a fixed dimensional linear vector that forms the input sequence for the transformer. This transformation converts a 224x224 image into a sequence of 196 patches. Each patch is represented by a 768-dimensional vector. The sequence of embedded patches is then passed into a stack of transformer encoder layers. Each layer consists of a multi-head self-attention mechanism followed by a feed-forward neural network. Here the layer normalization and residual connections are used to stabilize training and improve convergence. The attention mechanism allows the model to focus on different parts of the image simultaneously, capturing both local and global features. The DeiT model architecture is demonstrated in the figure 5 A special class token is prepended to the sequence of patch embeddings that is a

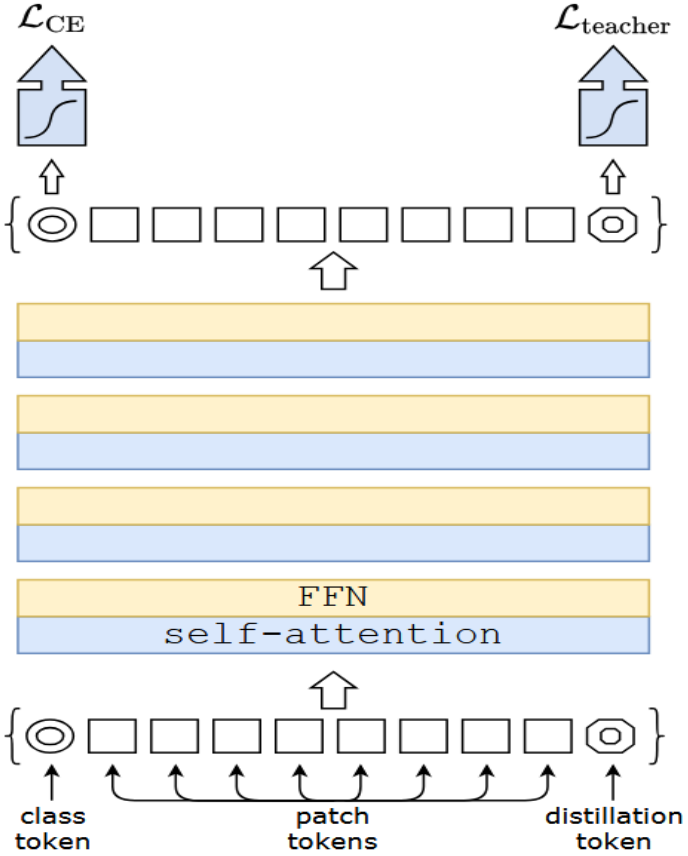

Figure 5: Data Efficient Image Transformer Network Architecture, (Touvron et al. (2021))

linear summary of the whole image. This token interacts with the other patches through the self-attention layers and is ultimately used for classification. The final hidden state corresponding to the class token is passed through a fully connected layer to produce the classification logits. This

output is then used to predict the class label of the input image. Data augmentation techniques such as random cropping, flipping, and color jittering are also used during pre-training as it can help to improve model generalization.

DeiT's key innovation is the use of knowledge distillation, where a smaller student model learns from a larger teacher model. With the help of this technique it enhances the data efficiency of the model. During training, the student model is helped by the soft labels produced by the teacher model, which provide more information than hard labels alone. After pre-training, the DeiT model is fine-tuned on the Galaxy10 dataset. Fine-tuning involves updating the model weights to adapt to the specific characteristics of galaxy images. Techniques such as learning rate scheduling and early stopping are also used in this research to optimize the fine-tuning process.

4.4 ResNet50

The ResNet50 architecture is a Convolutional neural network model in the field of deep learning. It is renowned for its exceptional performance in image classification tasks. It introduced an approach to training deep networks through the use of residual networks. This addressed the gradient degradation problem often encountered in very deep networks. ResNet50, short for Residual Network with 50 layers, consists of a series of convolutional layers, batch normalization layers, ReLU activation functions, and a unique feature known as identity mapping. This architecture helps the training of much deeper networks by allowing gradients to flow through shortcut connections directly to earlier layers, effectively mitigating the vanishing gradient problem.

The ResNet50 architecture is composed of 50 layers, including convolutional layers, pooling layers, and fully connected layers. The Figure 6 shows the ResNet50 network architecture. The architecture is divided into five stages, each with a different number of convolutional blocks. Each block contains multiple convolutional layers and makes use of identity mappings to create shortcut connections. The model starts with a 7x7 convolutional layer with 64 filters, followed

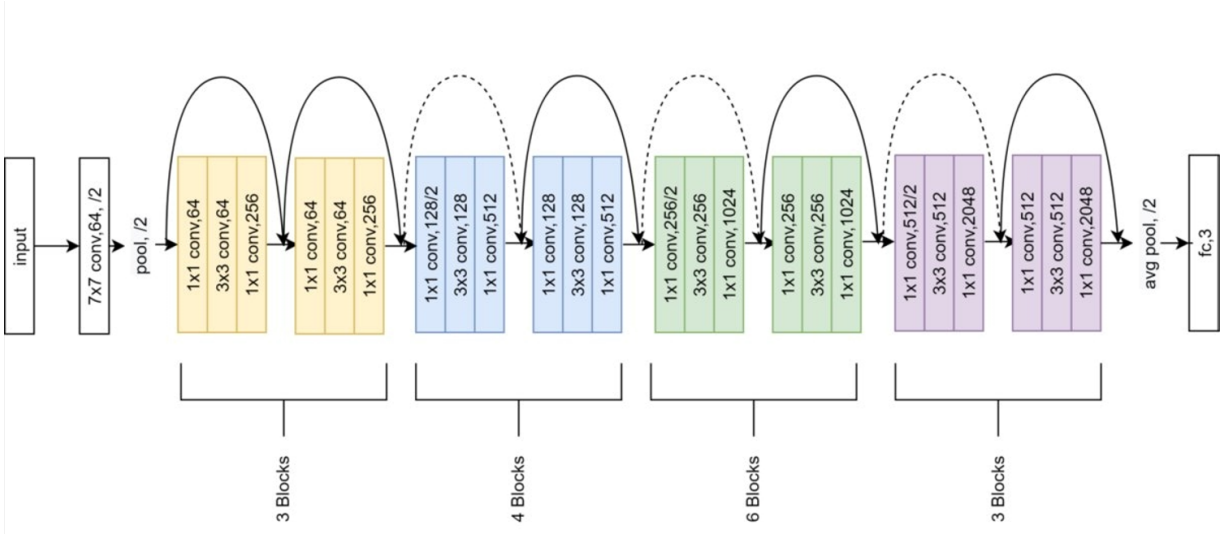

Figure 6: ResNet-50 Network Architecture, (Al-Humaidan and Prince (2021))

by a max pooling layer. This initial layer is responsible for extracting low-level features from the input image. The ResNet50 consists of residual blocks, each containing three convolutional layers with filter sizes of 1x1, 3x3, and 1x1. The first and last layers of each block are 1x1 convolutions that reduce and then restore dimensions, while the middle layer is a 3x3 convolution that processes the features.

After passing through the residual blocks, the feature maps are averaged via a global average pooling layer, which reduces the spatial dimensions to a single vector per filter. This vector is then fed into a fully connected layer with 1000 units (for ImageNet classification) or adjusted to match the number of classes in the specific task. ReLU activation and batch normalization are applied throughout the network to improve model convergence.

Following pre-training, the ResNet50 model is fine-tuned on the Galaxy10 dataset. Fine-tuning involves adjusting the model weights to adapt to the unique characteristics of galaxy images, which may include varying scales, orientations, and visual complexities. Techniques such as learning rate scheduling, early stopping, and data augmentation are employed during fine-tuning to optimize performance and prevent overfitting.

5 Implementation

In this project, four deep learning algorithms are used - ViT Base, Swin Transformer, DeiT Transformer and ResNet50 baseline. These models are evaluated based on their capability of classifying the different classes of the galaxy images present in the Galaxy10 Decals dataset. Each and every model uses the categorical cross entropy as the loss function. Mostly Adam Optimizer is used for the optimization of the models. One of the key aspect of implementation included selecting and tuning the hyperparameters of the models. The primary hyperparameters considered in this research are learning rate, batch size, number of epochs of training, weight decay and choice of optimizer. The learning rates for the models were set to a relatively low value to prevent the models from making large updates to the weights. The batch size was chosen as 8 as it was appropriate for the available GPU memory.A grid search approach was used in this research for testing various combinations of learning rates, batch sizes, and optimizer configurations. Adam optimizer was chosen for its ability to adapt learning rates for individual parameters.

During training, the cross-entropy loss function was used as it is better suited for multi-class classification tasks. One of the significant challenges faced during training was the long convergence time, particularly with Vision Transformers. To mitigate this challenge early stopping was implemented, such that training could be stopped if the validation loss did not improve after a certain number of epochs.

The training of Vision Transformers is computationally intensive. The implementation was carried out on a local machine equipped with an NVIDIA GeForce GTX 1650 GPU with 4 GB of memory, 16 GB of RAM, and a 256 GB Solid State Drive. The choice of hardware imposed certain limitations, particularly in terms of batch size and the number of epochs of model training over the galaxy dataset. Due to the computational resource constraint the vision transformers took significantly longer time to train when compared to the ResNet50 model.

The evaluation of the models was conducted using a carefully designed pipeline. The Galaxy10 DECals dataset was split into training, validation, and test sets in a 70:10:20 ratio. This split helped to reserve a portion of the data for unbiased evaluation. The code was written in Python 3.8.8 and multiple different packages and libraries were utilized, including Torch, Pytorch, Matplotlib, NumPy, tqdm, time, Sklearn, PIL, h5py and some others. The models make use of the GPU tensor cores with the help of PyTorch's data loaders.

6 Evaluation

As described in the research question, a primary target is to find out whether vision transformers are any better than traditional convolutional neural networks. In other words, we want to identify a transformer with efficient learning time and high accuracy. So that it can be determined whether or not Vision Transformers are worth considering for general-purpose image classification tasks. To make comparisons fair, a relatively complex image classification dataset

was employed. The galaxy 10 decals dataset has been divided into 10 classes, one for a morphological different galaxy. As there is class imbalance, the comparison needs to consider it as well.

6.1 Experiment 1 - Evaluation of Base Model ResNet50 and Other Vision Transformer Models *without* Data Augmentation

Analysis of the Precision, Recall, F1 Score, Accuracy and Weighted Average for the models that are trained over the data that does not incorporate data augmentation is detailed below in this experiment. The data is just resized into 224x224 pixels so that the models can take a standard image size as the input.

6.1.1 Evaluation Metrics - Precision

The precision scores for the four models are compared based on the classification of various galaxy types. The model comparisons are shown for precision in the Figure 7. The ViT base

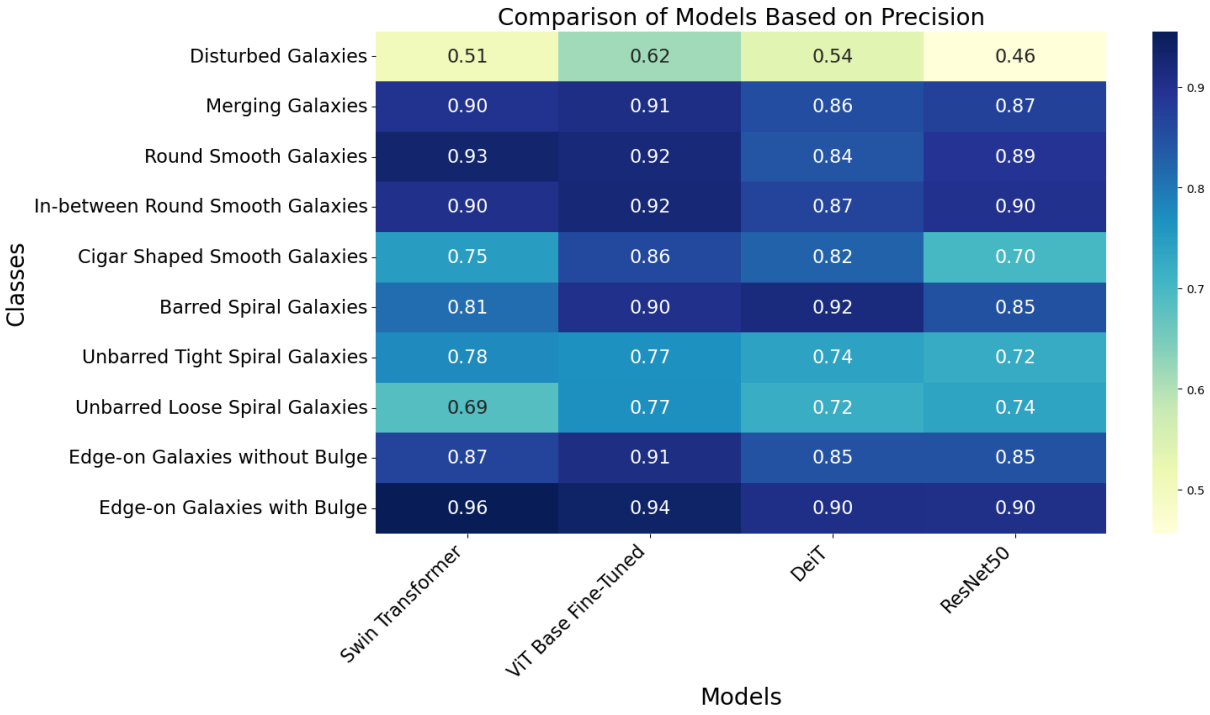

Figure 7: Precision Comparison for all models without Data Augmentation

Fine-Tuned model outperforms all the other models in this case. The peak precision noted for this model was for the "Edge-on Galaxies with Bulge" at around 94%. Swin Transformer model achieved an overall precision of 83%. The highest precision was for "Edge-on Galaxies with Bulge" at 96%, and the lowest for "Disturbed Galaxies" at 51%. DeiT Transformer: Achieved an overall precision of 81%. The highest precision was for "Barred Spiral Galaxies" at 92%, and the lowest was again for "Disturbed Galaxies" at 0.54. The baseline ResNet50 model achieved an overall precision of 81%. The highest precision achieved by this model was for "Edge-on Galaxies with Bulge" at 90%.

6.1.2 Evaluation Metrics - Recall

Upon checking the recall for the models, the Swin Transformer achieved an overall recall of 83%, performing best in the "Edge-on Galaxies without Bulge" category with a recall of 95%, while its recall for "Disturbed Galaxies" was 49%. The ViT Base Fine-Tuned model outperformed others with an overall recall of 86%, with a high recall of 96% for "Edge-on Galaxies with Bulge". The DeiT Transformer achieved an overall recall of 82%, excelling in "In-between Round Smooth Galaxies" with a recall of 97%. Finally the baseline ResNet50 model matched the DeiT Transformer with an overall recall of 82%. It performed the best in "Round Smooth Galaxies" (95%) and worst in "Disturbed Galaxies" (37%).

6.1.3 Evaluation Metrics - Accuracy Weighted Average

Accuracy reflects the proportion of true results among the total number of cases. In this study the Vit Base-Fine tuned model again achieved the highest overall accuracy and the weighted average as well. The model's weighted average precision, recall, and F1-Score, all are at 86%. This indicates a very stable and robust model for the research purpose. The Swin transformer's over accuracy and weighted average was at about 83%. In this experiment both the DeiT Transformer and ResNet50 models achieved an overall accuracy of 82% while the weighted average for both models were at 81%.

6.1.4 Evaluation Metrics - F1 - Score

The model comparisons based on the F1 score is shown in the figure 8. The Swin Transformer achieved an overall F1-Score of 83%, with its highest score for "Edge-on Galaxies with Bulge" at 93%. The ViT Base Fine-Tuned model led the evaluation with an overall F1-Score of 86%, achieving its highest score for "Edge-on Galaxies with Bulge" at 95%. Both the DeiT Transformer and ResNet50 had an overall F1-Score of 81%. As usual all the model performed poorly for the "Distributed Galaxies" class.

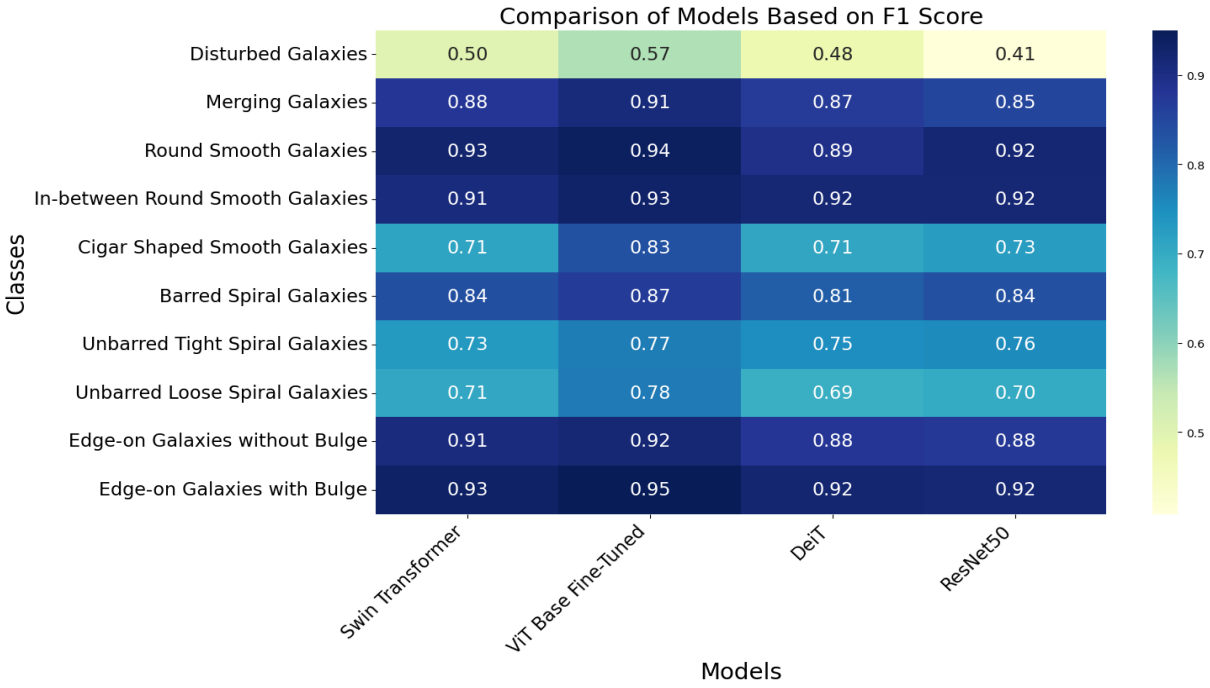

Figure 8: F1 Score Comparison for all models without Data Augmentation

6.2 Experiment 2 - Evaluation of Base Model ResNet50 and Other Vision Transformer Models *with* Data Augmentation

Analysis of the Evaluation of Baseline ResNet50 and Other Vision Transformers over the test dataset when the models are trained with data augmentation incorporated during training. There are multiple data augmentation included such as resizing, where the images are resized into a standard dimension of 224x224 pixels. Other augmentation such as random rotation of images into 90 degrees of rotation, horizontal and vertical flipping and random cropping as well. On top of this the images are also normalized to a specific range as this could speed up the convergence of the training process.

6.2.1 Evaluation Metrics - Precision

By the help of the precision metric the evaluation of the model checks its ability to correctly identify positive samples among the predicted positives. The Overall comparison of models for precision is shown in Figure 9.

In this research the Swin Transformer achieved a macro average precision of 85%, which shows a overall robust performance across different galaxy classes. The ViT Base Fine-Tuned model also performed well, with a macro average precision of 86%. It showed particularly high precision for categories like Edge-on Galaxies without Bulge (94%) and Edge-on Galaxies with Bulge (94%). The DeiT model had a macro average precision of 81%, with notable precision in categories such as Round Smooth Galaxies (89%) and Edge-on Galaxies with Bulge (92%). The Baseline ResNet50 model achieved a macro average precision of 81%. Its precision was high for classes like Round Smooth Galaxies (88%) and Edge-on Galaxies with Bulge (89%).

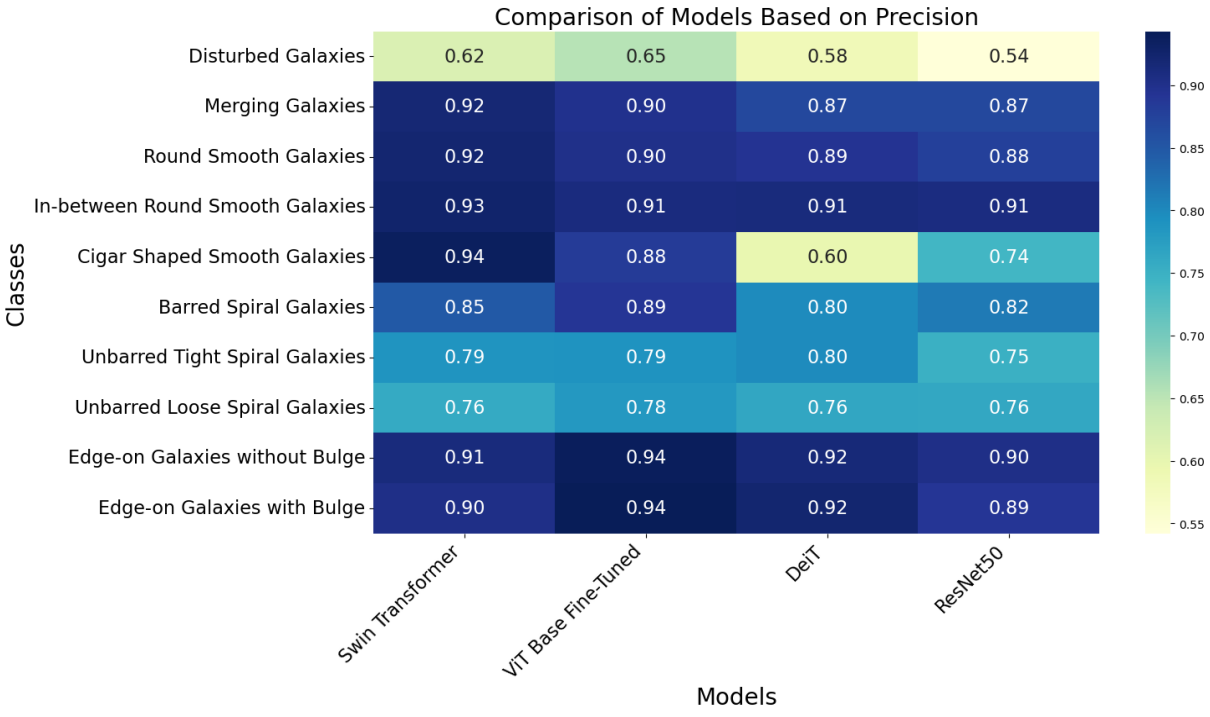

Figure 9: Precision Comparison for all models with Data Augmentation

6.2.2 Evaluation Metrics - Recall

The Swin Transformer model demonstrated a strong macro average recall of 82%. It performed really well in identifying Edge-on Galaxies with Bulge (96%) and Unbarred Loose Spiral Galaxies

(79%). The ViT Base Fine-Tuned model achieveEd an overall macro average recall of 85%. It showed excellent statistics in categories such as Round Smooth Galaxies (96%) and In-between Round Smooth Galaxies (96%). The DeiT model achieved a macro average recall of 83%, excelling in classes like Round Smooth Galaxies (96%) and In-between Round Smooth Galaxies (95%). The Baseline ResNet50 had a macro average recall of 81%. It showed strong recall for Round Smooth Galaxies (96%) and Edge-on Galaxies with Bulge (93%) over the test dataset. The DeiT model achieved a macro average recall of 83%, excelling in classes like Round Smooth Galaxies (96%) and In-between Round Smooth Galaxies (95%). The Baseline ResNet50 had a macro average recall of 81%. It showed strong recall for Round Smooth Galaxies (96%) and Edge-on Galaxies with Bulge (93%) over the test dataset.

6.2.3 Evaluation Metrics - F1 Score

The F1 score is the harmonic mean of precision and recall. It provides a single metric to evaluate the models' performance. The Overall comparison of models for F1 score is shown in Figure 10.

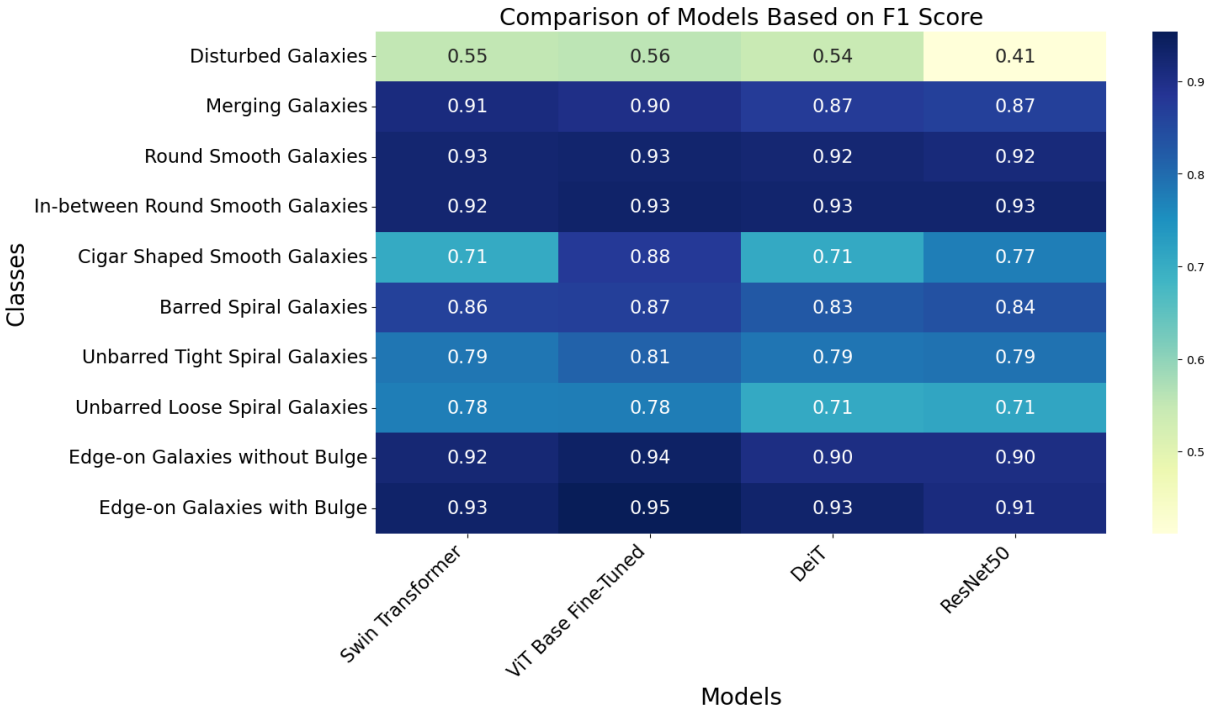

Figure 10: F1 Score Comparison for all models with Data Augmentation

The Swin Transformer showed a macro average F1 score of 83%. It was particularly effective in classes like Round Smooth Galaxies (93%) and Edge-on Galaxies with Bulge (93%). The ViT Base Fine-Tuned model achieved a macro average F1 score of 85%, with high F1 scores in categories like Round Smooth Galaxies (93%) and In-between Round Smooth Galaxies (93%). The DeiT model had a macro average F1 score of 81%, performing well in classes such as Round Smooth Galaxies (92%) and Edge-on Galaxies with Bulge (93%). The Baseline ResNet50 model achieved a macro average F1 score of 81%. Its F1 score was notable for Round Smooth Galaxies (92%) and Edge-on Galaxies with Bulge (91%).

6.2.4 Evaluation Metrics - Accuracy Weighted Average

Accuracy indicates the overall correctness of the models' predictions taking into account class imbalance. The metrics is used in this research to check how correctly the models categorize all

the images in test dataset to their respective classes. The Swin Transformer model achieved an accuracy of 86%, demonstrating a balanced performance across all galaxy classes. The weighted average that the model achieved was 85.57%. Now the ViT Base Fine-Tuned model had the highest accuracy at 87%, showcasing its capability to correctly classify a large majority of the galaxy images. The overall weighted average that the model achieved across all the classes was 86.32%. The DeiT model's accuracy was 84%, indicating solid performance but with some room for improvement. The weighted average across 10 classes was 83.30%. The Baseline ResNet50 model achieved an accuracy of 83%, performing well but slightly lower than the transformer-based models. Results of this section address research question one.

6.3 Experiment 3 - Evaluation of Base Model ResNet50 and other Vision Transformers based on training time

Upon checking the model training times it can be seen thet ResNet50 model takes significantly less time compared to the other Vision Transformer models. The ResNet50 baseline model take an average time of around 2 hour 20 minutes whereas the other vision transformer models take atleast 10 hours to train. The baseline model trained around 22 epochs with a convergence on the 17th epoch whereas the vision transformers are taking almost an hour for each epoch of the model training. ViT Base finetuned is the only transformer model that takes the least amount of time to train among all the vision transformers. ViT Base fine-tuned model takes approximately 19 minutes to train each epoch. The total training time for the model considering data augmentation was about 7 hours 31 minutes to train about 24 epochs while converging model in the 14th epoch. The table 1 shows the time consumed by the models to train for the stated number of epochs. The table notes down the training time of the models that train over the data that also has data augmentation incorporated as models performs better than the one that does not involve data augmentation. Results from this section are relevant to research question two.

Models	Time	Epochs
Swin Transformer	10:04:57	10
ViT Base Finetuned	07:31:08	24
DeiT	09:49:15	10
ResNet50	02:15:10	22

Table 1: Training Time by Model

6.4 Discussion

Once all the steps are done, the models are all trained twice. First the models are trained without any data augmentation and then second they are all trained over the dataset with data augmentation. Transfer learning was used in the study. A number of interesting findings follow:

- The models trained over the galaxy dataset when compared with each other it is evident that ViT Base Fine-Tuned model outperformed every other vision transformers and even the baseline model.

- The ViT base fine-tuned model achieved 86.16% over the dataset without any data augmentation, and about 86.70% over the dataset with data augmentation incorporated. Data augmentation thus only marginally improves the classification accuracy.
- This was the highest performance achieved by the model when compared with all other models. The model even had the least loss while testing as well where the test loss was about 0.4602 without augmentation and 0.4083 with augmentation.
- Other models performed quite high as well and they all had test accuracy above 80%. Swin Transformer model had the next best test accuracy of about 85.57%, and then DeiT and baseline ResNet both had around 83%. This clearly shows that the transformer models worked really well for the objective of this research.

In this research it was also noticed that the vision transformers did not demonstrate a significant rise in their performance when data augmentation was incorporated. This occurred as augmentation introduced more variety to the training data. However, it was seen that the loss during training for the training and validation set was marginally low for all the models. In terms of learning time, the following findings are of notice:

- Now, even though the baseline ResNet50 model had less but competitive performance, still upon checking the training time it had the least time consumption to train the models.
- ResNet50 generalized way faster than the vision transformers with about 2 hours and 15 minutes.
- Meanwhile, the fastest vision transformer model, ViT, took around 7 hours 31 minutes and the other vision transformer took more than 10 hours to converge into their best accuracy that they achieved.

In other words, ResNet50 was more than 5 hours faster than the best transformer model. The ViT model Here ViT base fine-tuned model took around 19 minutes approximately to train each epoch where as the other vision transformers took almost about an hour for each epoch. ResNet50 however only used around 5 minutes for each epoch training. Thus ResNet50 trains the faster when compared to all other models.

When compared with the baseline ResNet50 model, the vision transformers performed better than the Convolutional Neural Network model. This addresses the first research question and clearly indicates that vision transformer are worth researching in this field.

Along with this ViT Base finetuned model performs the best in terms of accurately classifying the galaxies among all other vision transformers. Swin Transformer comes next and then Data Efficient Image transformer followed the other models in terms of accuracy. This addresses the second research question of identifying the best vision transformer model.

6.5 Limitations

Despite the promising findings and advancements presented in this study there are several limitations that needs to be addressed. One of the primary limitations of this study is the reliance on the Galaxy10 DECals dataset, which is relatively small in size. Even though it provides crucial information on galaxy classes as it contains 10 classes still it does not fully capture the diversity and complexities in galaxy morphologies. There are bigger datasets explored in other research in this field like the study by (Lin et al.; 2022). Additionally, the dataset contans imbalance, where certain galaxy classes are underrepresented. The choice of using Vision Transformers (ViTs) over traditional Convolutional Neural Networks (CNNs) was driven by their ability to capture global dependencies. However, this study did not fully explore the potential of combining both approaches, such as using hybrid models that integrate CNNs with ViTs as done by

(Cao et al.; 2024). Besides this this study relies on a single dataset which is split into training, validation, and test sets which raises concern about potential bias in the splits. Although stratified sampling was used to make sure that each set maintained a similar class distribution still the inherent randomness in splitting the data could introduce bias affecting model performance. To address such issues rigorous cross-validation techniques could be implemented such as k-fold cross validation as done by (Reza; 2021).

Finally, one last crucial limitation faced during this research was the computational resource constraint. The study was conducted on a GPU with only 4GB of memory which doesn't allow a lot of tensor cores for the model training. This limitation could be avoided with the help of using better GPU such as NVIDIA A100 GPUs, as they can leverage about 40GB or more memory.

7 Conclusion and Future Work

In this research, the effectiveness of Vision Transformers in the morphological classification of galaxies was explored using the Galaxy10 DECals dataset. Three advanced transformer-based models—ViT Base, Swin Transformer, and DeiT Transformer—were compared among each other while considering the conventional ResNet50 model as the baseline. The primary objective was to evaluate how well these models were able to accurately classify galaxy morphologies. The research also helps to determine if Vision Transformers could outperform traditional CNNs or at least provide competitive results. The findings of this study are three fold:

1. ViT Base model achieved the highest accuracy at 86.7%, followed closely by the Swin Transformer and DeiT Transformer models for both the cases of model training i.e, with and without data augmentation. They all exhibited high precision and recall across the 10 galaxy classes.

2. The baseline model (ResNet50) performed really well, however, it lagged slightly behind the transformer-based models. These results demonstrate that Vision Transformers had superior ability to gain contextual information through self-attention mechanisms and they are highly effective for the task of galaxy morphology classification.

3. However in terms of training time ResNet50 showed significantly faster speed. It trained 22 epochs in about 2 hours 15 minutes while fastest vision transformer took 7 hour 31 minutes to train 24 epochs.

A systematic evaluation was performed. The study's evaluation metrics were recall, precision, accuracy, and F1 score. Thus, it provides a comprehensive assessment of the models' performances. The consistent performance of the transformer-based models shows their robustness and potential for detailed morphological analysis of galaxies. This research clearly presents the usefulness of Vision Transformers in astronomical image classification and suggests that they can significantly enhance the accuracy and efficiency of automated galaxy classification systems, and indeed be used for general purpose image classification tasks.

While the systematic evaluation of vision transformers has shown very promising results, there is scope for future work. For instance, the architecture of vision transformers could be adapted to perform faster. Or, a more balanced dataset could be used by incorporating additional galaxy datasets. The models can also be further fine-tuned using AutoML techniques. Besides this multiple models could be combined to form a ensemble model which would be able to leverage the strengths of multiple models and possibly perform better. One of the noble aspect from this research is the use of better computational resource. A better GPU could have more tensor cores and could be able to run more parallel computing which would allow the models to run faster. This could possibly lead to some interesting results that would allow further tuning

of the models in a faster way. By addressing these future possibilities the research can continue to enhance the effectiveness of automated galaxy morphology classification.

References

Abell, P. A., Allison, J., Anderson, S. F., Andrew, J. R., Angel, J. R. P., Armus, L., Arnett, D. and S.J. *et al* (2009). *LSST Science Book, Version 2.0.*
URL: *https://arxiv.org/pdf/0912.0201.pdf*

Ackermann, S., Schawinski, K., Zhang, C., Weigel, A. K. and Turp, M. D. (2018). Using transfer learning to detect galaxy mergers, *Monthly Notices of the Royal Astronomical Society* **479**(1): 415–425.
URL: *http://dx.doi.org/10.1093/mnras/sty1398*

Al-Humaidan, N. A. and Prince, M. (2021). A classification of arab ethnicity based on face image using deep learning approach, *IEEE Access* **9**: 50755–50766.
URL: *https://doi.org/10.1109/ACCESS.2021.3069022*

Applebaum, K. and Zhang, D. (2015). Classifying galaxy images through support vector machines, *2015 IEEE International Conference on Information Reuse and Integration*, pp. 357–363.
URL: *http://dx.doi.org/10.1109/IRI.2015.61*

Barchi, P., da Costa, F., Sautter, R., Rosa, R. and Carvalho, R. (2016). Improving galaxy morphology with machine learning.
URL: *http://dx.doi.org/10.6062/jcis.2016.07.03.0114*

Cao, J., Xu, T., Deng, Y., Deng, L., Yang, M., Liu, Z. and Zhou, W. (2024). Galaxy morphology classification based on Convolutional vision Transformer (CvT), **683**: A42.
URL: *http://dx.doi.org/10.1051/0004-6361/202348544*

Cavanagh, M. K., Bekki, K. and Groves, B. A. (2021). Morphological classification of galaxies with deep learning: comparing 3-way and 4-way CNNs, *Monthly Notices of the Royal Astronomical Society* **506**(1): 659–676.
URL: *https://doi.org/10.1093/mnras/stab1552*

de Vaucouleurs, G., de Vaucouleurs, A., Corwin, Herold G., J., Buta, R. J., Paturel, G. and Fouque, P. (1991). *Third Reference Catalogue of Bright Galaxies:Volume III.* New York: Springer Science+Business Meda, LLC.
URL: *https://doi.org/10.1007/978-1-4757-4363-0*

Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., Tuccillo, D. and Fischer, J. L. (2018). Improving galaxy morphologies for SDSS with Deep Learning, *Monthly Notices of the Royal Astronomical Society* **476**(3): 3661–3676.
URL: *https://doi.org/10.1093/mnras/sty338*

Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., Kaviraj, S., Fischer, J. L., Abbott, T. M. C. and Abdalla et al (2018). Transfer learning for galaxy morphology from one survey to another, *Monthly Notices of the Royal Astronomical Society* **484**(1): 93–100.
URL: *https://doi.org/10.1093/mnras/sty3497*

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale.
URL: *https://arxiv.org/abs/2010.11929*

Huertas-Company, M., Aguerri, J. A. L., Bernardi, M., Mei, S. and Sánchez Almeida, J. (2010). Revisiting the hubble sequence in the sdss dr7 spectroscopic sample: a publicly available bayesian automated classification, *Astronomy amp; Astrophysics* **525**: A157.
URL: *http://dx.doi.org/10.1051/0004-6361/201015735*

Lin, J. Y.-Y., Liao, S.-M., Huang, H.-J., Kuo, W.-T. and Ou, O. H.-M. (2022). Galaxy morphological classification with efficient vision transformer.
URL: *https://arxiv.org/abs/2110.01024*

Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M. J., Nichol, R. C., Szalay, A., Andreescu, D., Murray, P. and Vandenberg, J. (2008). Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey, *Monthly Notices of the Royal Astronomical Society* **389**(3): 1179–1189.
URL: *http://dx.doi.org/10.1111/j.1365-2966.2008.13689.x*

Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., Yuan, Z., Huang, Y., Sun, H., Gao, J., He, L. and Sun, L. (2024). Sora: A review on background, technology, limitations, and opportunities of large vision models.
URL: *https://arxiv.org/abs/2402.17177*

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows.
URL: *https://arxiv.org/abs/2103.14030*

Marin, M. A., Sucar, L. E., Gonzalez, J. A. and Díaz, R. (2013). A hierarchical model for morphological galaxy classification, *The Florida AI Research Society*.
URL: *https://api.semanticscholar.org/CorpusID:12189740*

Pearson, W. J., Wang, L. and Trayford, J. W. et al. (2019). Identifying galaxy mergers in observations and simulations with deep learning, *Astronomy & Astrophysics* **626**: A49.
URL: *https://doi.org/10.1051/0004-6361/201935355*

Reza, M. (2021). Galaxy morphology classification using automated machine learning, *Astronomy and Computing* **37**: 100492.
URL: *https://www.sciencedirect.com/science/article/pii/S2213133721000469*

Schneider, J., Stenning, D. C. and Elliott, L. T. (2023). Efficient galaxy classification through pretraining, *Frontiers in Astronomy and Space Sciences* **10**.
URL: *https://doi.org/10.3389/fspas.2023.1197358*

Shaiakhmetov, D., Mekuria, R. R., Isaev, R. and Unsal, F. (2021). Morphological classification of galaxies using spinalnet, *2021 16th (ICECCO)*, pp. 1–5.
URL: *https://doi.org/10.1109/ICECCO53203.2021.9663784*

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. and Jégou, H. (2021). Training data-efficient image transformers distillation through attention.
URL: *https://arxiv.org/abs/2012.12877*

York, D. G., Adelman, J., Anderson, Jr., J. E., Anderson, S. F., Annis, J. and et al, B. (2000). The Sloan Digital Sky Survey: Technical Summary, *The Astronomical Journal* **120**(3): 1579–1587.
URL: *http://dx.doi.org/10.1086/301513*