EXPLORING THE CAPABILITIES OF VISION TRANSFORMER MODELS FOR IMAGE-BASED GALAXY MORPHOLOGICAL CLASSIFICATION

ANDRAS TUU

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

126988

COMMITTEE

Mirella De Sisto
dr. Henry Brighton

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

May 19th, 2023

WORD COUNT

8074

ACKNOWLEDGMENTS

"We've always defined ourselves by the ability to overcome the impossible. And we count these moments. These moments when we dare to aim higher, to break barriers, to reach for the stars, to make the unknown known. We count these moments as our proudest achievements."

EXPLORING THE CAPABILITIES OF VISION TRANSFORMER MODELS FOR IMAGE-BASED GALAXY MORPHOLOGICAL CLASSIFICATION

ANDRAS TUU

Abstract

The amount of astronomical data collected by different observatories has been exponentially increasing in recent years. Classifying the astronomical objects captured through different means is a crucial part of processing astronomical data. The thesis is aiming to contribute to the field of galaxy classification by advancing accuracy and understanding the interplay between data quality, model performance, and observational advancements. The datasets used are Galaxy10SDSS, Galaxy10DECals, and a sample of the Galaxy10DECals dataset, with high redshift. The methodology consists of preprocessing the images, with Supervised and Unsupervised Wiener Deconvolution, Chan-Vese segmentation, and Non-Local Means denoising. For the multi-class classification of the preprocessed and original images, CNN and ViT models are used. The preprocessing of the images generally increased the models' performance. The best model in most cases was the ViT model, while for the high redshift sample, the CNN model performed better. The best preprocessing methods were Non-Local Means Denoising and Supervised Wiener Deconvolution. In the aspect of computational efficiency, the best preprocessing method was Supervised Wiener Deconvolution, while the ViT model was more efficient than the CNN.

0.1 *Source/Code/Ethics/Technology Statement*

Data Source: The (Galaxy10SDSS, Galaxy10DECals) have been acquired from the (https://github.com/henrysky/Galaxy10) through an online request. Work on this thesis did not involve collecting data from human participants or animals. All the figures belong to the author. Part of the CODE

has been adapted by the author from (https://github.com/amartinazzo/label-the-sky, licensed under a CC0 BY 4.0). The reused/adapted code fragments are clearly indicated in the notebook. In terms of writing, the author used assistance with the language of the paper. The author utilized the free tool provided by Grammarly for spell checking and grammar, and the generative language model GPT-3.5, in order to improve upon the author's original content. The code used in this thesis is not publicly available.

The following parts are mandatory to include in any document, thesis or article, that uses these datasets. Thus, the following part is mandatory to include when using the Sloan Digital Sky Survey (SDSS): Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS Web Site is http://www.sdss.org/. The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington.

Furthermore, the following part is mandatory to include when using the Galaxy10DECals dataset and repository: Copyright 2017-2023 Henry Leung Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions: The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software. THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND

1 INTRODUCTION

In recent years, there has been an exponential growth in the amount of data collected from astronomical telescopes worldwide. This increase has allowed for more accurate information gathering and classification of astronomical objects. Data science methods are fundamental for processing, classifying, and analyzing these vast amounts of data. Numerous systems exist to ensure the efficient storage and classification of such information. For instance, the Sloan Digital Sky Survey generates approximately 200 GB of data every night (Kollmeier et al., 2017).

One of the paramount challenges in the field of astronomy object classification is the expense associated with manually classifying data to train Machine Learning or Deep Learning models. Additionally, the quality of the image becomes a critical factor when identifying objects in astronomical images, as models tend to classify brighter images with more success.

This research paper explores the potential of Vision Transformers as an alternative to existing methods, especially when dealing with challenging-to-classify data. It delves into the capabilities of the self-attention mechanism proposed by Dosovitskiy et al. (2021) and its potential to improve the classification of fainter images. The primary focus of this research is on the classification of galaxies based on image data, hence it will not explore Machine Learning models. Additionally, it aims to investigate whether Vision Transformers can serve as a viable alternative to Convolutional Neural Network models, especially when considering harder-to-classify data.

In order to provide a structured approach to this study, the research revolves around the main research question:

1. How does the performance of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) compare when classifying astronomical objects?

    (a) Comparing the performance between two datasets, one being an older and imbalanced set with lower-quality images and

the other being a newer, more balanced set with higher-quality images?

(b) Considering factors like time and the computational power required, which solution emerges as the most efficient?

(c) Which methods yield the highest accuracy when classifying hard-to-classify images?

2 RELATED WORK

There are various ways that have been and are yet to be explored for astronomical object classification. The ones that are closely connected to the research are Machine Learning (ML) models, CNNs and Vision transformers, and Crowdsourcing efforts.

2.1 *Machine Learning Models*

Supervised learning methods, such as Support Vector Machines (SVMs) and Random Forests, have proven to be effective when applied to labeled datasets (Baron, 2019). Unsupervised techniques, such as clustering algorithms, have also shown potential, especially when dealing with vast amounts of unlabeled data. Ensemble methods, which leverage the power of multiple learning algorithms, offer another promising approach. The ML methods use various factors and information such as the Thuan-Gunn astronomic magnitude system (Thuan & Gunn, 1976), a photometric redshift system that is broadly used in classifying galaxies. The most recent articles and some older ones using ML are (Ball et al., 2006; Barchi et al., 2020; Mahabal et al., 2017). ML models for astronomical image classification have been explored for a long period.

2.2 *Crowdsourcing*

Initiatives to label astronomical data through crowdsourcing efforts also exist, with Galaxy Zoo (Lintott et al., 2008) as an example. This platform asks volunteers questions in a Decision Tree design to classify galaxies. The most common responses are used for the final classification. It's worth mentioning that labeling astronomical data is costly. This challenge is highlighted by Martinazzo et al. (2020), who emphasize that large amounts of training data are required to effectively train neural networks.

2.3 *Deep Learning Models*

Convolutional Neural Networks (CNNs) have become increasingly popular in the field of astronomical object classification and are considered one of the best models for image classification tasks. CNNs can learn the features from raw data, but training these models from scratch is computationally expensive. The limitation of training these models is the necessity of large amounts of labeled training data. And while large amounts of data are there, labeling them correctly is a limitation not yet solved. Deep Learning models for astronomical object classification have been explored in multiple studies such as (Ackermann et al., 2018; Chan & Stott, 2019; González et al., 2018). These articles are exploring transfer learning, models made for specific classifications. The article this research is based on is the one by Martinazzo et al. (2020), which is a comprehensive guide that explored the most popular models (Alexnet, VGG16, ResNet50, InceptionResNetV2, InceptionV3, DenseNet121), with Imagenet weights and with training from scratch, optimizers, hyperparameters.

2.4 *Vision Transformer Models*

Transformer models are one of the most significant recent advances in the field, and these models can also be applied for different image segmentation, classification, and other computer vision tasks. The model was proposed by Dosovitskiy et al. (2021). ViTs apply the transformer architecture originally proposed by Vaswani et al. (2017) to image data, allowing the models to capture long-range dependencies across the entire image. The ViT models, as for all the computer vision field, are a great innovation in the context of astronomical image classification as well. These models are able to focus on the most informative parts of an image through the self-attention mechanism. The ViT models are evolving rapidly; however, compared to ML and CNN models, their application in astronomical image classification has not been as extensively researched. The research of Lin et al. (2022) uses a Linformer-based model in order to classify galaxies. This model did not use the image pre-processing of Misra et al. (2018) and used the Linformer ViT model, which achieved an accuracy of 80.55%. This knowledge gap will be explored in this research. While exploring the ViT/B16 Vision Transformer model's performance, which was proposed by Wu et al. (2020).

3 METHODOLOGY

3.1 *Dataset Description*

The two datasets used in this research were retrieved from astroNN (Walmsley et al., 2021). These datasets have been widely used by other researchers for classification. The Galaxy 10 SDSS dataset contains images from the Sloan Digital Sky Survey (Kollmeier et al., 2017). The Galaxy10 SDSS dataset was created with Galaxy Zoo using approximately 270,000 SDSS galaxy images, of which approximately 22,000 were classified into 10 broad categories by volunteers. The classification process employed a Decision tree model, and only pictures where more than 55% of votes agreed on the class of the object were used. The authors chose this threshold based on the performance of a neural network model (Cifar10), which is an alternative for classification. The original images, with dimensions of 424x424 pixels, were cropped to a 207x207 pixel region centered on the galaxies as part of preprocessing. They were then downscaled three times using bilinear interpolation to a final image size of 69x69 pixels, in order to enhance computational efficiency. Since most of the image is just black space, this cuts out the galaxy that the research focuses on. Also, it is important to mention that the classes in this dataset are not balanced, thus this research primarily focuses on the other dataset, Galaxy10 Decals. However, it is important to test the pre-processing steps and models on the older smaller resolution dataset to have some comparison between newer images, captured by more state-of-the-art observatories, and the older images.

Table 1 below presents the classes and the number of images classified in each.

Class	(Number of Images)	Morphology
0	(3461 images)	Disk, Face-on, No Spiral
1	(6997 images)	Smooth, Completely round
2	(6292 images)	Smooth, in-between round
3	(394 images)	Smooth, Cigar shaped
4	(1534 images)	Disk, Edge-on, Rounded Bulge
5	(17 images)	Disk, Edge-on, Boxy Bulge
6	(589 images)	Disk, Edge-on, No Bulge
7	(1121 images)	Disk, Face-on, Tight Spiral
8	(906 images)	Disk, Face-on, Medium Spiral
9	(519 images)	Disk, Face-on, Loose Spiral

Table 1: Galaxy10 SDSS dataset (21785 images) Source: (Kollmeier et al., 2017; Lintott et al., 2008)

The second dataset is also classified by individual volunteers from Galaxy Zoo on the DESI Legacy Imaging Surveys (DECals)(Dey et al., 2019), resulting in approximately 441,000 unique galaxies that were filtered more rigorously. Approximately 18,000 of these images were classified into 10 broad categories, with a few small amendments to make each class more distinct. The Edge-on Disk with Boxy Bulge category, which originally only had 17 images, was removed from both datasets. This dataset contains 17736 256x256 images colored using g, r, and z bandwidth. The dataset also contains ans, ra, dec, redshift, and pxscale in units of an arcsecond per pixel. Out of the variables mentioned here, the redshift variable holds significant value. Redshift is a key cosmological parameter used in astronomy to estimate the distance of galaxies from Earth. It works on the principle of the Doppler Effect, where the observed color of light changes depending on the motion of the source with respect to the observer. In the case of galaxies, a higher redshift implies that the galaxy is moving away from us at a faster rate, which in turn typically indicates a greater distance. There is a correlation between redshift and the distance between the object and the observatory capturing the image, based on Hubble's law. The redshift data in this dataset provides an indirect measure of how far each galaxy is and can be particularly useful when determining the hard-to-classify images as a category. Building upon the data provided, images for this analysis will be strategically selected based on their distance from Earth, which makes them inherently more challenging to classify. As the redshift value increases, indicating greater distance, the light we receive from these galaxies becomes fainter and the overall image quality diminishes. This is primarily because light has to traverse a larger portion of the universe, leading to greater attenuation and distortion due to various cosmological phenomena. Consequently, the images of such high redshift galaxies tend to be more challenging to decipher and classify accurately. Table 2 below presents the classes and the number of images classified in each

Class	(Number of Images)	Morphology
0	(1081 images)	Disturbed Galaxies
1	(1853 images)	Merging Galaxies
2	(2645 images)	Round Smooth Galaxies
3	(2027 images)	In-between Round Smooth Galaxies
4	(334 images)	Cigar Shaped Smooth Galaxies
5	(2043 images)	Barred Spiral Galaxies
6	(1829 images)	Unbarred Tight Spiral Galaxies
7	(2628 images)	Unbarred Loose Spiral Galaxies
8	(1423 images)	Edge-on Galaxies without Bulge
9	(1873 images)	Edge-on Galaxies with Bulge

Table 2: Galaxy10 DECals dataset (17736 images) Source: (Dey et al., 2019; Walmsley et al., 2022)

3.2 *Image preprocessing*

CNNs and ViTs can benefit from these preprocessing steps because both types of models learn to make predictions by identifying patterns in the input data. If the input data is cleaner and more focused on the aspects of the image that are relevant to the prediction task (because of segmentation, deconvolution, or denoising), it should be easier for the models to identify the correct patterns and make accurate predictions.

Furthermore, in the case of ViTs specifically, as they split the image into patches and treat each patch as a token, clear and noise-free images could lead to better representation of these tokens, hence better performance. In this research, multiple image processing methods will be used, based on the most recent and comprehensive scientific article on astronomy image preprocessing: Advanced Image Processing for Astronomical Images by Misra et al. (2018):

- Non-local Means Noise Removal
- Supervised and Unsupervised Wiener-Hunt Deconvolution
- Chan-Vese Segmentation
- Data augmentation

3.2.1 *Non-local Means Noise Removal*

Noise removal is an essential process, with the main aim being to smoothen the image and maintain as much information as possible. Traditional linear noise removal techniques like Gaussian smoothing or median filtering tend to blur the edges and reduce the fine details in images. These methods are computationally efficient and remove the noise, but they often fail to preserve the edges, resulting in image blurring. Finding the optimal solution that balances preserving the original image and reducing the noise is crucial.

The Non-Local Means algorithm is a non-linear model that replaces the intensity value of the target pixel with the average of an array of intensities from other pixels. It compares small regions centered around the other pixels to the region centered around the target pixel, and performs averaging when the two regions are very similar. This process helps preserve the details and texture present in the image while reducing noise. This method is particularly effective for images with repetitive textures or patterns, such as the Galaxy datasets. The σ value, which represents the estimated noise standard deviation, is taken from the research conducted by Misra et al. (2018) with $\sigma = 0.3567791543719770$.

Figure 1 and Figure 2 show the results of the NLMeans denoising applied to an example image from the Galaxy10DECals dataset.

Figure 1: Original Image

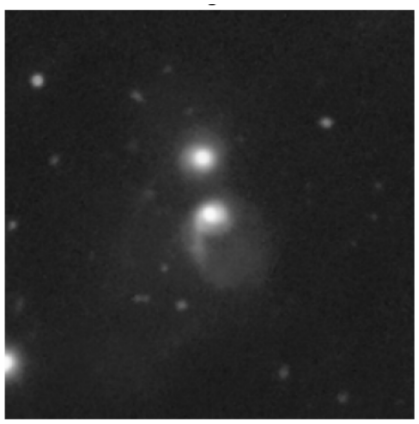

Figure 2: NLMeans denoised image

Furthermore, the Non-Local Means method is a non-linear method, which is better for preserving features in the image but is less efficient in reducing the noise in comparison to the Wiener-Hunt deconvolution. Also, the Wiener-Hunt deconvolution, being a linear model, offers improved computational efficiency, especially in its supervised version. The processing power and computational efficiency discussed in the Image Preprocessing section (Section 3.2) are of utmost importance in the field of astronomical image classification.

3.2.2 *Wiener-Hunt deconvolution*

This model is commonly utilized for denoising images. Both the supervised and unsupervised versions of this linear model have been tested. As discussed in the Non-local Means Noise Removal section (Section 3.2.1), these models remove noise quickly and efficiently but are less effective in preserving edges.

The model incorporates the Point Spread Function (PSF), which measures the spreading of a simple point of light in the image. It also applies penalization to high frequencies, typically considered as noise, in an effort to minimize them in the images. The model aims to strike a balance between preserving the image and reducing noise as much as possible.

The two approaches to the Wiener-Hunt deconvolution: Supervised and Unsupervised.

1. *Supervised Wiener-Hunt Deconvolution*: In the supervised approach, the spectral characteristics of both the original signal and the noise are estimated. In this method, a denoising parameter of 0.711 is

implemented based on extensive experimentation. This value is chosen to strike a balance between effective denoising without excessive blurring that could result in the loss of important image features. The Wiener filter utilizes these estimated parameters to minimize the mean square error between the original and deconvolved signal. This approach strengthens the parts of the image where the signal is stronger than the noise while weakening the parts where the noise dominates.

2. *Unsupervised Wiener-Hunt Deconvolution*: In the unsupervised approach, the algorithm itself learns the estimation for the noise using the Gibbs Sampler. This self-learning model seeks to find different parameters that optimize the trade-off between noise reduction and preserving edges and other important details. However, it is important to note that this process is computationally expensive. Figure 3, Figure 4, and Figure 5 shows the results of the Supervised and Unsupervised Wiener-Hunt deconvolution applied to an example image from the Galaxy10DECals dataset.

Figure 3: Original Image

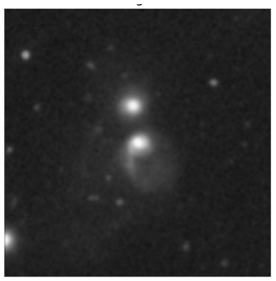

Figure 4: Supervised Wiener image

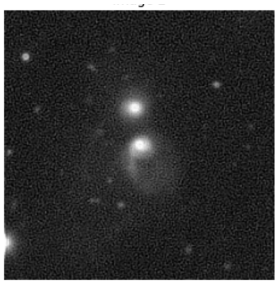

Figure 5: Unsupervised Wiener image

3.2.3 *Chan-Vese Segmentation*

The Chan-Vese segmentation model, explored in the research conducted by Misra et al. (2018), is a valuable tool for astronomical image processing. This model is particularly useful for object segmentation, which plays a crucial role in image-processing tasks. In astronomical images, including those in the Galaxy datasets, the presence of noise and the lack of clear boundaries pose challenges for traditional edge detection-based segmentation models. The Chan-Vese segmentation model, on the other hand, performs well with images that lack well-defined boundaries.

The model iteratively minimizes an energy function, which consists of weighted sums of intensity differences from the average value inside and outside the segmented region, as well as a term based on the length of the edge of the segmented region. In the research conducted by Misra et al. (2018), the following parameter values were used: $\mu = 0.5$, $\lambda_1 = 1$, and $\lambda_2 = 2$. These specific parameter values were chosen based on the irregular distribution of objects and their background in the dataset. Figure 6 and Figure 7 shows the results of the Chan-Vese segmentation applied to an example image from the Galaxy10DECals dataset.

Figure 6: Original Image

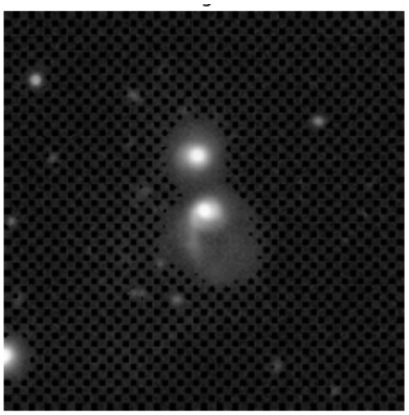

Figure 7: Chan-Vese processed image

3.2.4 *Data Augmentation*

Data augmentation artificially increases the size of the training dataset by applying random transformations to the images, such as rotation, width and height shifts, zooming, and horizontal flipping. This helps the model generalize better and also is a step that reduces overfitting. The methods, that worked best with the model were implemented, with the optimal parameters. For this, previous personal experience was used, with Deep

Learning projects, where data augmentation was implemented in a neural network model.

For all the image processing steps mentioned above, with the exception of Data Augmentation. The images were transformed to grayscale since it is a requirement for these image-processing models. The author is aware that with the transformation, there could have been some data loss in the color channels. This will be addressed further in the Discussion 5 part.

3.3 *Classification models*

3.3.1 *Convolution Neural Network*

The model used as the baseline is a VGG16 CNN model. The model was the generally best-performing one in the research of Martinazzo et al. (2020). The hyperparameter tuning, the optimizer settings, and regularization are all based on the research of Martinazzo et al. (2020). However, in this research, a new state-of-the-art optimizer named "Lion" was explored, but this did not improve the results. The Lion optimizer proposed by Chen et al. (2023) increases computational efficiency, compared to the Adam optimizer that all the models in this research use, and can be explored more in future research. This model has been originally trained on the ImageNet dataset which includes millions of images. By leveraging a pre-trained model, the baseline model can benefit from the learned features from this extensive dataset. This is particularly important considering that the datasets used in this research are relatively small for training CNN and ViT models from scratch.

The VGG16 model serves as the base model. The output of this model is passed through several additional layers, implemented based on the model used by Martinazzo et al. (2020). The GitHub repository of this research can be accessed here: https://github.com/amartinazzo/label-the-sky.

The model includes the following layers, parameters, and optimizers. A GlobalAveragePooling2D layer, which performs spatial averaging of the features. This helps the model to reduce the total number of parameters, making it more compact and less prone to overfitting. A Dense layer with a LeakyReLU activation function and 1024 nodes. The LeakyReLU function is a variant of the ReLU (Rectified Linear Unit) function, that allows small negative values when the input is less than zero, mitigating the frequently arising problem with ReLU where the neurons become inactive and no longer contribute to the learning of the model. In the model, there is also a maximum norm constraint applied to the weights, creating an upper bound for the values of each neuron. This also helps the model with overfitting.

A Dropout layer is added to the model, that randomly sets a fraction of the inputs to 0 during training, at each update. This also helps to prevent overfitting. The last layer is a Dense layer with 10 nodes (corresponding to the 10 classes of galaxies) and a softmax activation function for the classification, the outputs of this layer are the probabilities of the image belonging to each of the 10 classes specified.

During training, the model uses an Adam optimizer with an exponential decay learning rate schedule, which reduces the learning rate gradually over the training process. This enables the model to "learn" faster at the start of the training and only make smaller updates later on, thus helping the model converge to an optimal set of weights.

The model training is conducted in three stages:

At the start, the layers of the VGG16 model are frozen, and only the top layers are trained for 10 epochs. After the first 10 epochs, the last 7 layers of the VGG16 model are unfrozen, allowing them to be fine-tuned. Lastly, the entire model is trained for up to 200 epochs, with the following callback functions. The two callback functions implemented are Earlystopping and ModelCheckpoint. The Earlystopping stops the training of the model if the validation loss does not improve for 10 consecutive epochs. This helps with preventing the model from overfitting and also makes the model more computationally efficient. The ModelCheckpoint is used to save the model with the best performance on the validation set, with this the best model can be retrieved. The data is split into train, validation, and test sets. The labels are one-hot encoded, and the images are normalized before training and fitting the model.

3.3.2 *Vision Transformer*

The advanced model implemented is the Vision Transformer (ViT) model with the B16 configuration. This model was also pre-trained on the ImageNet dataset. Similarly, to the CNN model, this allows the model to utilize parameters learned from a much larger dataset, and it is only fine-tuned on the galaxy dataset. Training Vision Transformer models from scratch is a very computationally expensive process. To train and fit the model, which includes 17736 images with dimensions 224,224,3, the code required around 49 gigabytes of RAM. As discussed in the Introduction (see 3.4.1), the scarcity of data is not a limitation in this process.

The output of the ViT model is fed through multiple additional layers since the pre-trained model outputs 1000 classes, the same as the number of classes on ImageNet.

The model incorporates a Dropout layer for regularization, which aids in mitigating overfitting, and also enables the model to have better general-

izability. The final layer is a Dense layer with 10 nodes, corresponding to the 10 classes of galaxies. This layer uses a softmax activation, as the model is performing multiclass classification. Furthermore, L2 regularization is applied to the weights of the Dense layer, reducing the complexity of the models and overfitting the training data.

As for the fine-tuning process, the model incorporates an ADAM optimizer, although other optimizers such as RADAM, WADAM, and the LION optimizer proposed by Chen et al. (2023) were also explored. While the model converged faster with the Lion optimizer and took less time to run, the validation accuracy of the model decreased. The learning rate scheduler reduces the learning rate over the epochs, similar to the CNN model. This approach aids the model to make larger updates initially, and as the model achieves higher validation accuracy and the validation loss decreases, the learning rate is reduced to allow the model to find the optimal solution faster.

The model's two callback functions are the same as in the CNN model: Earlystopping and Model checkpoint. The data is also divided into training, validation, and test sets, the labels are one-hot encoded, and the images are normalized before training and fitting the model.

3.4 *Experimental Design*

3.4.1 *Introduction*

Understanding and classifying galaxies is a fundamental task in astrophysics. With the availability of large-scale galaxy datasets, such as Galaxy10SDSS and Galaxy10DECals, there is an opportunity to employ advanced computational techniques to improve classification accuracy. This thesis presents a comprehensive experimental design that combines image preprocessing, data augmentation, and deep learning models to improve galaxy classification. The experimental design including the Transformation and processing of the images and the classification of the images are visible in Figure 28.

3.4.2 *Image Preprocessing*

The design process begins with image preprocessing. During this research, four different preprocessing methods were applied to enhance the image quality and extract relevant features. These methods are supervised and unsupervised Wiener-Hunt deconvolution, Non-Local Means denoising, and Chan-Vese segmentation. These techniques aim to reduce noise, improve image resolution, and enhance the visibility of important structures within the galaxies. It is important to mention, that these image prepro-

cessing steps were applied one by one, during the research there were no combinations of masks or preprocessing steps explored. By assessing the results of the models on these varied image sets, we can measure the impact of preprocessing techniques on various evaluation metrics. For the balanced Galaxy10DECals dataset, classification accuracy, along with other evaluation methods, is employed. These additional methods include confusion matrices, ROC and AUC curves, precision, recall, F1-score, and support. Conversely, due to the imbalance in the SDSS dataset, classification accuracy is not considered as a reliable metric, and thus, the focus is directed towards the remaining evaluation measures.

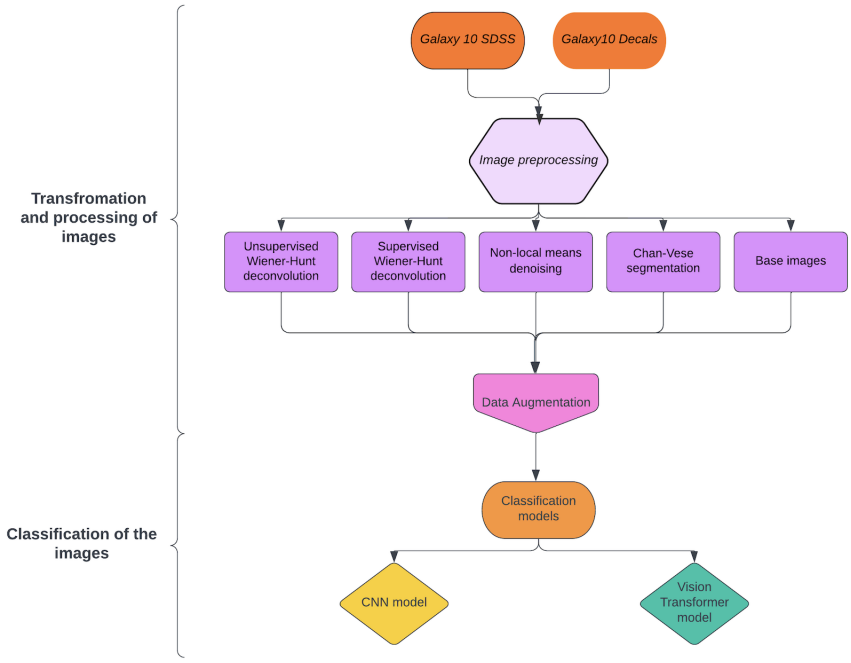

Figure 8: Experimental Design

3.4.3 *Data Augmentation and Train-Validation-Test Split*

To augment the dataset and address issues related to limited labeled data availability, data augmentation techniques have been applied. These techniques include random rotation, flipping, scaling, and cropping of the images. The augmented dataset is then divided into training, validation, and test sets.

3.4.4 *Normalization and Deep Learning Models*

By normalizing the pixel values, the models are more robust to differences in illumination and contrast among the images. The VGG16 pre-trained CNN network and the ViT-B16 (Vision Transformer) model are used for the classification of the galaxies. Both models have been used for image classification before, and the models based on the research of (Lin et al., 2022; Martinazzo et al., 2020; Wu et al., 2020) were used. By comparing the performance of these models on the different image sets, it becomes possible to evaluate their suitability for galaxy classification.

3.4.5 *Google Colab and Hardware Specifications*

For all the preprocessing training fitting and evaluation tasks, the Premium+ version of Google Colab was used. This way it was possible to utilize the V100 GPU with 16GB of VRAM, 51GB of RAM, and 250GB of storage. Which has significantly accelerated the training of the models, and made it possible to have 18 different models.

3.4.6 *Low Light Emission Analysis*

One aspect of the study focuses on identifying low light emission images based on redshift. The top 10% of images were selected, with the highest redshift values in each category, thus not making the Galaxy10DECals dataset imbalanced. This way also makes a subset of challenging images. The images only have the redshift values in the Galaxy10 DECals dataset, thus for the classification of the hard-to-classify images, there are only 10 models used, the CNN and the ViT architecture, and all the preprocessing image sets, along with the original images. By analyzing the classification results for these subsets, an insight can be gained into the relationship between redshift and galaxy morphological classification.

3.4.7 *Transfer Learning*

All models used in this research utilize Transfer Learning and are pre-trained on the ImageNet dataset. The models are only finetuned on the two datasets. With Transfer Learning, the models are using pre-trained weights, utilizing the training on the much larger Imagenet dataset. The transfer learning process for the CNN models involved freezing the initial layers to retain learned features and fine-tuning the last 7 layers to adapt to the Galaxy10SDSS and Galaxy10DECals datasets. This approach helped capture important features and meaningful representations from the galaxy images, potentially enhancing classification accuracy. Additionally, investigating the performance of these models on both older (Galaxy10SDSS)

and newer datasets (Galaxy10DECals) is also important. To see whether improved imaging technology, allowed for the models to perform better.

3.4.8 *Libraries used*

For the models and preprocessing of the images, the following packages were used:

- tensorflow (Martín Abadi et al., 2015) for building the models, layers, constraints, optimization, callbacks, and data augmentation.
- keras (Chollet et al., 2015) for the Vision Transformer model and also, for the parts mentioned already with TensorFlow.
- numpy (Harris et al., 2020) for transformations in arrays.
- openCV (Bradski, 2000) and scikit-image (Van der Walt et al., 2014) for transforming and preprocessing the images.
- Matplotlib (Hunter, 2007) for making the graphs of the evaluation metrics.
- hdf5 (Collette, 2013) for reading in the datasets and saving the best models.
- tqdm (da Costa-Luis, 2019) for measuring the progress of image processing

4 RESULTS

4.1 *Galaxy10DECals Dataset*

The results for the Galaxy10 DECals dataset are promising, and this dataset is particularly relevant for future research. The results of different models and metrics, which are used for the evaluation of the different models are accuracy, and macro averaged Precision, Recall, and F1 score. These metrics are calculated for every model in the Galaxy10 DECals dataset, these metrics can be seen in Table 3 below. The evaluation metrics show the effectiveness of the combinations of different preprocessing methods and classification models. As for the overall highest-performing combination of classification model and preprocessing in the DECALS dataset. The best-performing model is the Vision Transformer (ViT) model, with Non-Local Means denoising as the preprocessing method. This model performed the best in terms of accuracy and has the best balance across the other metrics out of these models. The model has an accuracy of 0.87, macro average

precision of 0.86, macro average recall of 0.84, and macro average F1 score of 0.85.

Table 3: Accuracy and macro average metrics for the DECALS dataset

Model	Preprocessing	Accuracy	Macro Avg		
			Precision	Recall	F1
CNN	OG	0.82	0.81	0.8	0.8
	CHANV	0.83	0.81	0.8	0.8
	NLM	0.85	0.84	0.83	0.83
	SUP	**0.86**	**0.84**	**0.84**	**0.84**
	UNSUP	0.79	0.77	0.77	0.76
ViT	OG	0.8	0.78	0.8	0.78
	CHANV	0.81	0.81	0.76	0.76
	NLM	**0.87**	**0.86**	**0.84**	**0.85**
	SUP	0.83	0.82	0.8	0.81
	UNSUP	0.82	0.8	0.8	0.8

4.1.1 *Comparison of CNN and ViT*

The Convolutional Neural Network (CNN) model performed better across all metrics on the original images compared to the Vision Transformer (ViT) model, especially when paired with supervised Wiener deconvolution preprocessing. This CNN model achieved an accuracy of 0.86, along with high precision (0.83), recall (0.83), and an F1 score of 0.84. However, it was slightly outperformed by the best-performing ViT model utilizing Non-Local Means preprocessing.

The CNN model showed particularly subpar performance with unsupervised Wiener deconvolution, achieving an accuracy of 0.79. In contrast, the ViT model's lowest accuracy was 0.8, observed when using both the original and Chan-Vese segmentation preprocessed images.

It's also noteworthy that the ViT model demonstrated more consistent performance across different preprocessing methods, suggesting that it is more robust compared to the CNN model. Robustness and generalizability are crucial aspects of these models, as the objective of this thesis and previous research is to contribute towards fully automating the galaxy classification process.

4.1.2 *Preprocessing Effects*

The preprocessing of the images was performed based on the research conducted by Misra et al. (2018). The preprocessing plays a crucial role in the performance of the classification models. The best-performing models for both the CNN and ViT architecture utilized two preprocessing methods, namely the Supervised Wiener deconvolution and Non-Local Means denoising. Notably, the CNN model's performance improved when using the Supervised Wiener Deconvolution method, while the ViT model achieved the best results with the Non-Local Means denoising. Also, the CNN model performed very well with the Non-Local Means denoising as well, while the second best performance of the Vision Transformer model is while using the Supervised Wiener Deconvolution for preprocessing.

The Unsupervised Wiener deconvolution method yielded lower performance compared to the supervised method across both models, which suggests that leveraging prior knowledge in the preprocessing stage is beneficial for these types of astronomical images. While, with the Supervised Wiener deconvolution, the model used the parameters mentioned by Misra et al. (2018), the Unsupervised model should be able to find the best parameters.

Supervised Wiener deconvolution uses a known point spread function, to recover the original image, this has helped the model better recognize features in the image data. Non-local Means denoising, on the other hand, is a powerful method that can reduce noise while preserving the structure of the image, this preprocessing technique is Non-Linear, thus helping with the preservation of the edges.

4.1.3 *Best performing models*

In the following part, the two best-performing models are going to be compared. Namely the Non-Local Means denoised Vision Transformer and the Supervised Wiener Deconvolution CNN model. As it is visible in Figure 9 and Figure 10 below, the ROC curves are very similar between the two models, with the ViT model outperforming the CNN model slightly in some areas, with the classification of Disturbed Galaxies and Merging Galaxies. It is visible on the ROC curves, that the models are performing well across all categories, except for Disturbed Galaxies.

The Classification report can be seen in Figure 11 and Figure 12. The models have good performance, based on accuracy and F1 score across all categories except for Disturbed Galaxies, where both models struggle. With the exception of Disturbed Galaxies, the models are generally well-rounded and perform well for both the positive and negative classes. Classes, where the CNN model performs better, are Round Smooth Galaxies, In-between

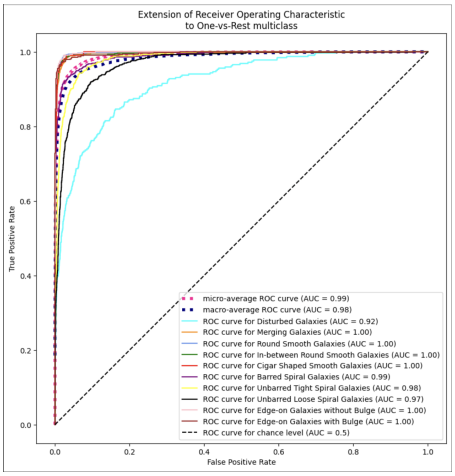

Figure 9: Non-local Means denoising ViT ROC-AUC curve

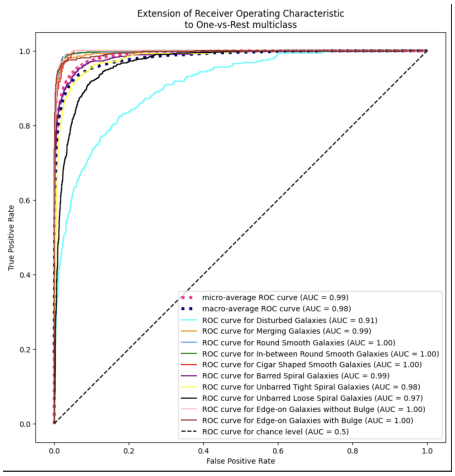

Figure 10: Supervised Wiener Deconvolution CNN ROC-AUC curve

Classification Report:

	precision	recall	f1-score	support
Disturbed Galaxies	0.74	0.40	0.52	319
Merging Galaxies	0.90	0.93	0.92	558
Round Smooth Galaxies	0.90	0.98	0.94	788
In-between Round Smooth Galaxies	0.90	0.96	0.93	584
Cigar Shaped Smooth Galaxies	0.86	0.77	0.81	108
Barred Spiral Galaxies	0.88	0.89	0.88	650
Unbarred Tight Spiral Galaxies	0.82	0.80	0.81	545
Unbarred Loose Spiral Galaxies	0.79	0.81	0.80	796
Edge-on Galaxies without Bulge	0.93	0.94	0.94	433
Edge-on Galaxies with Bulge	0.91	0.96	0.93	540
accuracy			0.87	5321
macro avg	0.86	0.84	0.85	5321
weighted avg	0.87	0.87	0.87	5321

Figure 11: Non-local Means denoising ViT Metrics

Classification Report:

	precision	recall	f1-score	support
Disturbed Galaxies	0.57	0.48	0.52	319
Merging Galaxies	0.93	0.89	0.91	558
Round Smooth Galaxies	0.94	0.94	0.94	788
In-between Round Smooth Galaxies	0.90	0.96	0.93	584
Cigar Shaped Smooth Galaxies	0.79	0.84	0.82	108
Barred Spiral Galaxies	0.88	0.86	0.87	650
Unbarred Tight Spiral Galaxies	0.88	0.74	0.80	545
Unbarred Loose Spiral Galaxies	0.72	0.84	0.77	796
Edge-on Galaxies without Bulge	0.91	0.92	0.91	433
Edge-on Galaxies with Bulge	0.93	0.94	0.93	540
accuracy			0.86	5321
macro avg	0.84	0.84	0.84	5321
weighted avg	0.86	0.86	0.86	5321

Figure 12: Supervised Wiener Deconvolution CNN Metrics

Round Smooth Galaxies, Edge-on Galaxies with Bulge, and Edge-on Galaxies without Bulge with higher precision, recall, and F1-score for these classes. The classes where the Vision Transformer model performs better are: Barred Spiral Galaxies, with higher precision, and Disturbed Galaxies with much higher precision, which is the largest difference between the two models across all metrics.

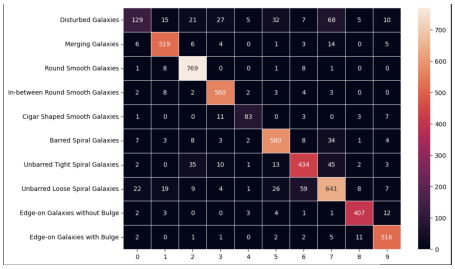

Figure 13: Non-local Means denoising ViT Confusion Matrix

Figure 14: Supervised Wiener Deconvolution CNN Confusion Matrix

The confusion matrices in Figure 13 and Figure 14 reinforce the low precision of the CNN model for Disturbed Galaxies, the model classifies more of these galaxies correctly, however, it produces many False Positives, a type 1. error, thus having low precision. Also, it is visible that both models have struggled with the classification of the Unbarred Loose Spiral galaxies, with the CNN model classifying more correctly, but producing a large number of False Positives.

Overall both models perform generally well across all categories, but for the difficult-to-classify images and categories, the Vision Transformer model performs better. In terms of computational efficiency, as discussed in the Image preprocessing section 3.2, the Non-Local Means denoising method is computationally demanding. For the dataset consisting of 17,736 images, denoising the pictures using the Non-Local Means method took approximately one hour, even with the assistance of a V100 GPU accelerator. In contrast, the Supervised Wiener Deconvolution took roughly six minutes on the same GPU accelerator.

Regarding the training time of the CNN and ViT models, they are comparable across all datasets, with the ViT models requiring about 10% less time for training. Both models generally perform well across all categories, but the Vision Transformer model shows superior performance when dealing with images and categories that are challenging to classify.

4.2 *High Redshift dataset*

The high redshift dataset is a sample from the Galaxy10 DECals dataset, that is created by extracting the top 10% of the images, with the corresponding labels based on the redshift values, as discussed in the Dataset Description 3.1. There is a correlation between redshift and the distance between the object and the observatory capturing the image, based on Hubble's law. The models performed very well on the Redshift dataset, with accuracy ranging from 0.77 to 0.90, these metrics can be seen in Table 4 below. In terms of preprocessing methods, the Non-Local Means denoising method and the Supervised Wiener Deconvolution, similar to the whole DECals dataset, seemed to be the most effective. Leading to the highest performance in both the Convolutional Neural Network (CNN) and Vision Transformer (ViT) models.

4.2.1 *Comparison of CNN and ViT*

The CNN model, when applied to the most distant, less light-emitting images, outperformed the ViT model. The best CNN model achieved an accuracy of 0.90 using Non-Local Means preprocessing. Conversely, the

Table 4: Accuracy and macro average metrics for the REDSHIFT dataset

Model	Preprocessing	Accuracy	Macro Avg		
			Precision	Recall	F1
CNN	OG	0.82	0.8	0.8	0.79
	CHANV	0.87	0.85	0.84	0.83
	NLM	**0.9**	**0.9**	**0.87**	**0.87**
	SUP	0.87	0.85	0.84	0.83
	UNSUP	0.85	0.84	0.83	0.82
ViT	OG	0.77	0.76	0.74	0.72
	CHANV	0.79	0.78	0.73	0.73
	NLM	0.86	0.85	0.84	0.84
	SUP	**0.87**	**0.88**	**0.86**	**0.86**
	UNSUP	0.84	0.84	0.82	0.83

best ViT model achieved an accuracy of 0.87 when Supervised Wiener Deconvolution preprocessing was implemented.

These results suggest that CNN's convolutional approach, which processes local features in the image in a hierarchical manner, is better suited to this specific part of the Galaxy10 DECals dataset. The ViT's self-attention mechanism appears to be less beneficial for these classification tasks.

4.2.2 *Preprocessing Effects*

Across both model types, the Non-Local Means denoising showed high performance, implying that noise reduction is crucial for images of the furthest galaxies. The Supervised Wiener Deconvolution preprocessing method also resulted in high performance across both models, with the best-performing Vision Transformer model utilizing this preprocessing method. Conversely, using the images unprocessed or with the Unsupervised Wiener Deconvolution typically resulted in lower performance. The application of parameters suggested by Misra et al. (2018) improved model performance.

4.2.3 *Best performing models*

The Classification reports, visible in Figure 17 and Figure 18, tell us, that the best-performing model for this task on the Redshift dataset is the CNN model with Non-Local Means denoising preprocessing. This model achieved the highest accuracy and F1 score. Among the ViT models, the one using Supervised Wiener Deconvolution preprocessing demonstrated

the best performance, albeit with slightly lower accuracy and F1 score compared to the CNN model.

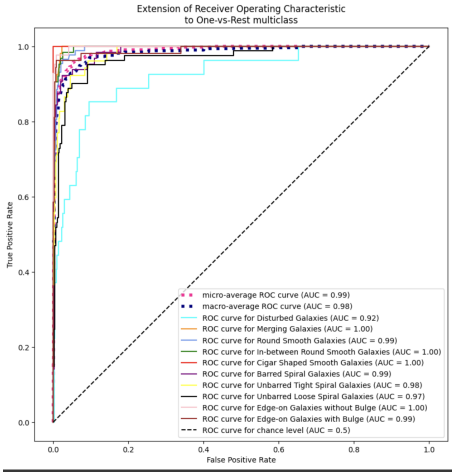

Figure 15: Supervised Wiener Deconvolution ViT ROC-AUC curve

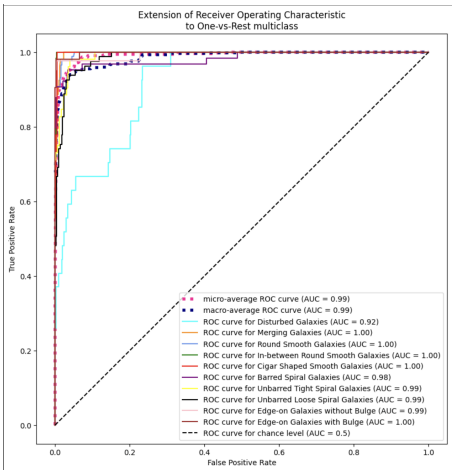

Figure 16: Non-local Means denoising CNN ROC-AUC curve

Classification Report:

	precision	recall	f1-score	support
Disturbed Galaxies	0.69	0.41	0.51	27
Merging Galaxies	0.85	0.98	0.91	53
Round Smooth Galaxies	0.86	0.97	0.91	87
In-between Round Smooth Galaxies	0.87	0.94	0.90	63
Cigar Shaped Smooth Galaxies	1.00	1.00	1.00	9
Barred Spiral Galaxies	0.92	0.84	0.88	64
Unbarred Tight Spiral Galaxies	0.88	0.71	0.79	52
Unbarred Loose Spiral Galaxies	0.81	0.84	0.82	81
Edge-on Galaxies without Bulge	1.00	0.93	0.96	42
Edge-on Galaxies with Bulge	0.93	0.96	0.94	53
accuracy			0.87	531
macro avg	0.88	0.86	0.86	531
weighted avg	0.87	0.87	0.87	531

Figure 17: Supervised Wiener Deconvolution ViT Metrics

Classification Report:

	precision	recall	f1-score	support
Disturbed Galaxies	0.90	0.33	0.49	27
Merging Galaxies	0.88	0.98	0.93	53
Round Smooth Galaxies	0.87	0.98	0.92	87
In-between Round Smooth Galaxies	0.93	1.00	0.96	63
Cigar Shaped Smooth Galaxies	0.75	1.00	0.86	9
Barred Spiral Galaxies	0.98	0.89	0.93	64
Unbarred Tight Spiral Galaxies	0.97	0.69	0.81	52
Unbarred Loose Spiral Galaxies	0.80	0.94	0.86	81
Edge-on Galaxies without Bulge	0.97	0.93	0.95	42
Edge-on Galaxies with Bulge	0.96	0.98	0.97	53
accuracy			0.90	531
macro avg	0.90	0.87	0.87	531
weighted avg	0.91	0.90	0.89	531

Figure 18: Non-local Means denoising CNN Metrics

The results are visible in Figure 15 and Figure 16. The models have good performance based on accuracy, F1 score, and the ROC-AUC curves, across all categories. The exception from this is the category Disturbed Galaxies, where both models struggle, similarly to the Galaxy 10 DECals dataset. However there is a big difference compared, to the last model, the CNN model in this case has much better precision than for the whole DECals dataset.

Similarly, to the whole dataset, with the exception of Disturbed Galaxies, the models are generally well-rounded and perform well for both the positive and negative classes. Classes, where the CNN model performs better, are Round Smooth Galaxies, Edge-on Galaxies with Bulge, Unbarred Tight Spiral Galaxies, Unbarred Loose Spiral Galaxies, and Disturbed Galaxies with much higher precision, which is the largest difference between the two models across all metrics. The classes where the Vision Transformer

model performs better are Edge-on Galaxies without Bulge, Barred Spiral Galaxies with higher precision, recall, and F1-score for these classes.

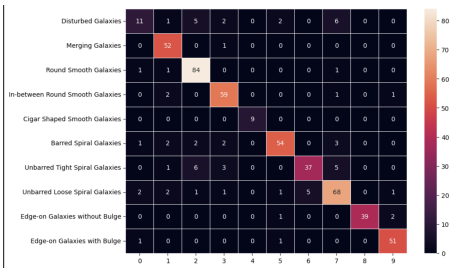

Figure 19: Supervised Wiener Deconvolution ViT Confusion Matrix

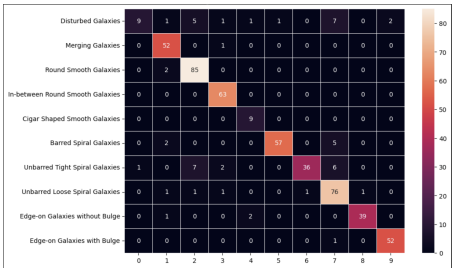

Figure 20: Non-local Means denoising CNN Confusion Matrix

The confusion matrices in Figure 19 and Figure 20 show the big change in terms of precision for the CNN model for Disturbed Galaxies, the model for this dataset classifies less of these galaxies correctly, it still produces some False Positives, but is mostly correct. Also, it is visible that both models have struggled with the classification of the Unbarred Loose Spiral Galaxies, with the CNN model classifying more correctly, but producing a larger number of False Positives.

For this specific part of the dataset, the CNN model performs better, across almost all categories, while the ViT model still performs adequately. Furthermore, as for the question of computational efficiency, as discussed in the 3.2, the Non-Local Means denoising method is computationally around ten times more expensive in comparison to the Supervised Wiener Deconvolution.

4.3 *Galaxy10SDSS dataset*

The models applied to the SDSS dataset, which is imbalanced, showcased a wide range of macro average precision, recall, and F1 values, ranging from 0.68 to 0.87, as it is visible in **??** It is important to mention that the accuracy measure is not a reliable metric in this case due to the nature of this dataset.

4.3.1 *Comparison of CNN and ViT*

Both the CNN and ViT models performed well in classifying the SDSS dataset. However, the ViT model slightly outperformed the CNN model, achieving the highest macro average precision of 0.87 using Non-local Means denoising as preprocessing compared to the best CNN model reaching 0.75 using the Chan-Vese segmentation preprocessing.

Table 5: Macro average metrics for the SDSS dataset (Accuracy not relevant due to imbalanced dataset)

Model	Preprocessing	Accuracy	Macro Avg		
			Precision	Recall	F1
CNN	OG	0.87	0.73	0.73	0.73
	CHANV	**0.86**	**0.75**	**0.72**	**0.73**
	NLM	0.86	0.75	0.71	0.73
	SUP	0.86	0.72	0.68	0.70
	UNSUP	0.85	0.70	0.68	0.69
ViT	OG	0.88	0.77	0.77	0.77
	CHANV	0.87	0.75	0.75	0.75
	NLM	**0.88**	**0.87**	**0.77**	**0.79**
	SUP	0.88	0.76	0.75	0.75
	UNSUP	0.87	0.75	0.74	0.74

4.3.2 *Preprocessing Effects*

The Non-local Means denoising preprocessing method consistently resulted in higher scores across both model architectures. This indicates that reducing noise is essential for the SDSS dataset as well. Conversely, to the previous results, the Supervised Wiener Deconvolution preprocessing method, underperformed the Non-local Means denoising preprocessing method but still performed better than other models.

4.3.3 *Best-performing models*

The ViT model with Non-local Means Denoising preprocessing achieved the highest macro average precision, recall, and F1 scores, making it the best-performing model. The CNN model with Chan-Vese preprocessing is the best-performing CNN model, demonstrating an adequate performance with a good balance between precision and recall.

It is visible in Figure 23, Figure 24, Figure 25, and Figure 26 as well, that the Galaxy10SDSS dataset is heavily imbalanced, the Non-Local Means denoising ViT model performed much better than the Chan-Vese segmentation. The better performance of the ViT model could be the result of a number of factors. ViT models compared to CNN models are more robust to the changes in resolution, by design. The ViT models can extract features regardless of their position or scale, while CNN models are very sensitive to the locality and scale of the features. Thus, ViT models this way are more suitable for lower resolution images, where important fea-

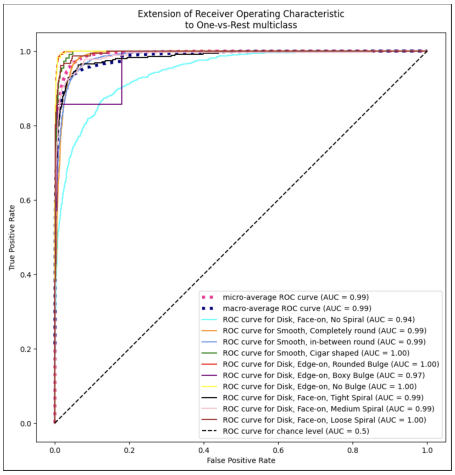

Figure 21: Non-local Means denoising ViT ROC-AUC curve

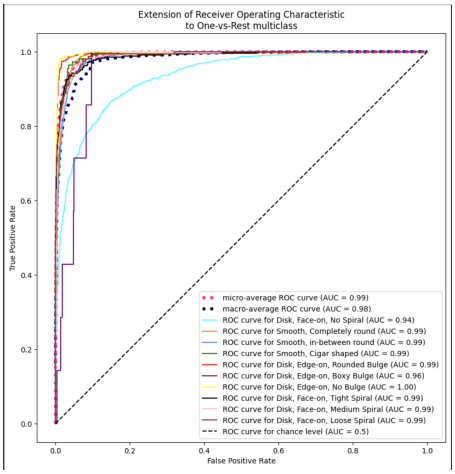

Figure 22: Chan-Vese Segmentation CNN ROC-AUC curve

Classification Report:	precision	recall	f1-score	support
Disk, Face-on, No Spiral	0.79	0.69	0.74	1033
Smooth, Completely round	0.92	0.94	0.93	2128
Smooth, in-between round	0.89	0.93	0.91	1863
Smooth, Cigar shaped	0.79	0.77	0.78	109
Disk, Edge-on, Rounded Bulge	0.91	0.97	0.94	457
Disk, Edge-on, Boxy Bulge	1.00	0.14	0.25	7
Disk, Edge-on, No Bulge	0.91	0.90	0.90	173
Disk, Face-on, Tight Spiral	0.76	0.83	0.79	345
Disk, Face-on, Medium Spiral	0.83	0.76	0.80	269
Disk, Face-on, Loose Spiral	0.91	0.81	0.86	152
accuracy			0.88	6536
macro avg	0.87	0.77	0.79	6536
weighted avg	0.88	0.88	0.88	6536

Figure 23: Non-local Means denoising ViT Metrics

Classification Report:	precision	recall	f1-score	support
Disk, Face-on, No Spiral	0.74	0.68	0.71	1033
Smooth, Completely round	0.92	0.94	0.93	2128
Smooth, in-between round	0.88	0.93	0.91	1863
Smooth, Cigar shaped	0.73	0.64	0.68	189
Disk, Edge-on, Rounded Bulge	0.92	0.88	0.90	457
Disk, Edge-on, Boxy Bulge	0.00	0.00	0.00	7
Disk, Edge-on, No Bulge	0.79	0.94	0.85	173
Disk, Face-on, Tight Spiral	0.86	0.76	0.81	345
Disk, Face-on, Medium Spiral	0.78	0.73	0.75	269
Disk, Face-on, Loose Spiral	0.84	0.72	0.78	152
accuracy			0.86	6536
macro avg	0.75	0.72	0.73	6536
weighted avg	0.86	0.86	0.86	6536

Figure 24: Chan-Vese Segmentation CNN Metrics

tures may be very subtle or spread across the image. While working with astronomy images, understanding the spatial relationships between distant parts of an image is essential. For instance, recognizing faint structures surrounding galaxies requires learning information from the entire image. The self-attention mechanism of the ViT architecture allows the model to grasp the global context. Furthermore, older images, such as the images in the Galaxy10SDSS dataset might contain complex patterns due to the technology and conditions at the time of capture. The self-attention mechanism of ViT enables the model to capture complex patterns better, as it learns to focus on the most informative parts of the image.

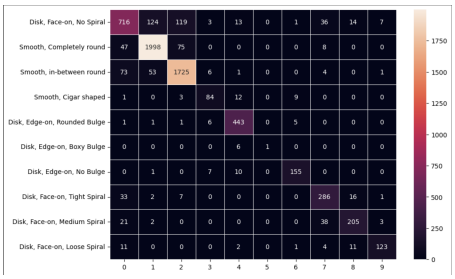

Figure 25: Non-local Means denoising ViT Confusion Matrix

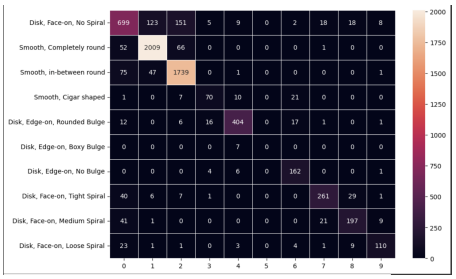

Figure 26: Chan-Vese Segmentation CNN Confusion Matrix

5 DISCUSSION

5.1 *Summary and Discussion of the Results*

The performance of the models on the three datasets were promising, both the Vision Transformer and CNN models performed well with different preprocessing methods, which are very important for the models, creating large improvements.

For the Galaxy10DECALS dataset, the models performed robustly, with accuracy ranging from 0.79 to 0.87. In terms of preprocessing techniques, the Non-Local Means denoising method produced the best performance in both the Convolutional Neural Network and Vision Transformer models.

As for the Redshift dataset, where the selection was based on the highest redshift values from the data in the Galaxy10DECals dataset. The CNN model utilizing Non-Local Means denoising preprocessing had the best accuracy of 0.90 in the whole research. This suggests that there is no connection between the distance of the galaxies and the difficulty of classification.

This can come down to a number of reasons. Firstly, the distance and speed of light. These galaxies can be classified easier, since they are observed as they were in the distant past, due to the time it takes for their light to reach us over a large distance. Furthermore, these galaxies might have more distinct, identifiable features because they represent earlier stages in the evolution of the universe, thus making them easier to classify rather than harder. Moreover, another explanation is that there could be cosmological factors at play. Galaxies at high redshifts are often experiencing rapid star formation, which could result in more distinct, recognizable features. The truly hard-to-classify galaxies are the ones that have shapes that are difficult to classify, for example, all the models have struggled with the classification of Disturbed Galaxies, where the Vision Transformer model outperformed the CNN models.

In the case of the SDSS dataset, which includes older and imbalanced data. The ViT models generally outperformed the CNN models. This could be attributed to ViT's ability to handle lower-resolution images and capture long-range dependencies.

The CNN models were performing similarly to ViT models on the DECALS and especially the Redshift datasets. In the case of the DECals dataset the CNN model with the Unsupervised Wiener Deconvolution, is much more computationally efficient than the best performing Non-local Means denoising ViT model. However, on the DECals dataset, the performance was similar, for the hard-to-classify images, such as Disturbed Galaxies the Vision Transformer model performed significantly better.

5.2 *Comparison to the Literature*

The preprocessing steps proposed by Misra et al. (2018), were essential for this research. The preprocessing methods played a significant role in model performance, with Non-Local Means denoising and Supervised Wiener Deconvolution, most of the times performing the best across all datasets. In most cases the models performed better, with the preprocessed images, thus image preprocessing is extremely relevant for galaxy morphological classification. Also, the best-performing image preprocessing methods are for image denoising, while trying to find the optimal solution between preserving the important features and reducing noise. The CNN models, proposed by Martinazzo et al. (2020) also performed very well, and played a crucial role, in creating a baseline model, to which the Vision Transformer models can be compared. Lastly, the Vision Transformer model first explored for galaxy morphological classification by Lin et al. (2022), performed well for image classification, this research explored this topic further, including the image preprocessing and a different Vision Transformer model.

5.3 *Scientific and Societal Impact*

The amount of astronomical data collected by different observatories has been exponentially increasing in recent years. Classifying the astronomical objects captured through different means is a crucial part of processing astronomical data. For the classification of astronomical objects, there have been Machine learning approaches on non-image data, crowd-sourcing efforts, and using images with CNN and now Vision Transformer models. The amount of data captured by the SDSS every night is around 200 GB. It is not feasible to classify data of this magnitude manually, thus exploring and creating automated approaches are of high importance. The thesis is aiming to contribute to the field of galaxy classification by advancing accuracy and understanding the interplay between data quality, model performance, and observational advancements.

5.4 *Limitations and future research*

During the conduct of this research, the most significant limitations were computing power, storage space, and RAM, even while utilizing Google Colab GPU acceleration. Moreover, due to these computing power limitations, the ViT model could not be fine-tuned using keras-tuner. Consequently, the hyperparameters in use are based on previous research and experimentation. This model could potentially be improved with keras-tuner.

Furthermore, there are new advancements in the rapidly evolving field of Vision Transformer models, such as Modelsoups (Wortsman et al., 2022), which is a method to combine the best-performing models from different ViT architectures, thereby achieving better performance. Another limitation, also mentioned by Martinazzo et al. (2020), is the scarcity of labeled data for training the models.

During the preprocessing phase, the images had to be converted to grayscale in order to perform the image preprocessing steps. In this process, there may have been data loss. If different image preprocessing methods can be applied to colored images while achieving the same denoising results, the models' performance could potentially be improved.

Lastly, during training, different optimizers and optimizer settings were explored, such as RADAM, WADAM, and the most recent advancements in the field of optimizers, including the new Lion optimizer proposed by Chen et al. (2023). This area could also be further explored. While the performance of the models did not increase when using the Lion optimizer, it was found to be computationally much more efficient than the ADAM optimizer

6 CONCLUSION

RQ1: How does the performance of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) compare to each other when classifying astronomical objects? In the results section, all the models are visible, generally, the ViT models outperform the CNN models, with the optimal image preprocessing methods applied. Also, Vision Transformer models are more robust and improved generalizability.

SRQ1.a: Comparing the performance between two datasets, one being an older and imbalanced set with lower quality images and the other being a newer, more balanced set with higher-quality images? On the older Galaxy10SDSS dataset, the Vision Transformer models outperformed the CNN models, while using the optimal image preprocessing technique. While on the newer Galaxy10DECals dataset, the CNN models performed similarly. For the high redshift dataset, the CNN models however outperformed the Vision Transformer models.

SRQ1.b: Considering factors like time and the computational power required, which solution emerges as the most efficient? Generally, the best-performing model is the Vision Transformer model, however, in certain cases, the CNN model performed better. For example, on the Galaxy10DECals dataset, with the Supervised Wiener Deconvolution preprocessing method. The training of a ViT model is

around 10% faster than the CNN model. Moreover, image preprocessing takes longer and is much more computationally expensive. The fastest preprocessing is the Supervised Wiener Deconvolution, while the Non-Linear Means denoising takes ten times longer. The Chan-Vese segmentation and Unsupervised Wiener Deconvolution take approximately twenty times longer than the Supervised Wiener Deconvolution.

RQ1.c: Which methods yield the highest accuracy when classifying hard-to-classify images? For the hard-to-classify images, that are in the Disturbed Galaxies category in the Galaxy10DECals dataset, the Vision Transformer models outperform the Convolutional Neural Network models.

REFERENCES

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

Ball, N. M., Brunner, R. J., Myers, A. D., & Tcheng, D. (2006). Robust Machine Learning Applied to Astronomical Data Sets. I. Star-Galaxy Classification of the Sloan Digital Sky Survey DR3 Using Decision Trees. *The Astrophysical Journal*, *650*(1), 497–509. https://doi.org/10.1086/507440

Barchi, P. H., de Carvalho, R. R., Rosa, R. R., Sautter, R., Soares-Santos, M., Marques, B. A. D., Clua, E., Gonçalves, T. S., de Sá-Freitas, C., & Moura, T. C. (2020). Machine and Deep Learning Applied to Galaxy Morphology – A Comparative Study [arXiv:1901.07047 [astro-ph]]. *Astronomy and Computing*, *30*, 100334. https://doi.org/10.1016/j.ascom.2019.100334

Baron, D. (2019). Machine Learning in Astronomy: A practical overview. https://doi.org/10.48550/ARXIV.1904.07248

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Chan, M. C., & Stott, J. P. (2019). Deep-CEE i: Fishing for galaxy clusters with deep neural nets. *Monthly Notices of the Royal Astronomical Society*, *490*(4), 5770–5787. https://doi.org/10.1093/mnras/stz2936

Chen, X., Liang, C., Huang, D., Real, E., Wang, K., Liu, Y., Pham, H., Dong, X., Luong, T., Hsieh, C.-J., Lu, Y., & Le, Q. V. (2023). Symbolic discovery of optimization algorithms.

Chollet, F., et al. (2015). Keras.

Collette, A. (2013). *Python and hdf5*. O'Reilly.

da Costa-Luis, C. (2019). Tqdm: A fast, extensible progress meter for python and cli. *Journal of Open Source Software*, *4*, 1277. https://doi.org/10.21105/joss.01277

Dey, A., Schlegel, D. J., Lang, D., Blum, R., Burleigh, K., Fan, X., Findlay, J. R., Finkbeiner, D., Herrera, D., Juneau, S., Landriau, M., Levi, M., McGreer, I., Meisner, A., Myers, A. D., Moustakas, J., Nugent, P., Patej, A., Schlafly, E. F., … Zhou, Z. (2019). Overview of the DESI Legacy Imaging Surveys., *157*(5), Article 168, 168. https://doi.org/10.3847/1538-3881/ab089d

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale [arXiv:2010.11929 [cs]]. https://doi.org/10.48550/arXiv.2010.11929

González, R. E., Muñoz, R. P., & Hernández, C. A. (2018). Galaxy detection and identification using deep learning and data augmentation.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Kollmeier, J. A., Zasowski, G., Rix, H.-W., Johns, M., Anderson, S. F., Drory, N., Johnson, J. A., Pogge, R. W., Bird, J. C., Blanc, G. A., Brownstein, J. R., Crane, J. D., De Lee, N. M., Klaene, M. A., Kreckel, K., MacDonald, N., Merloni, A., Ness, M. K., O'Brien, T., … van Saders, J. L. (2017). SDSS-V: Pioneering Panoptic Spectroscopy [arXiv:1711.03234 [astro-ph]]. https://doi.org/10.48550/arXiv.1711.03234

Lin, J. Y.-Y., Liao, S.-M., Huang, H.-J., Kuo, W.-T., & Ou, O. H.-M. (2022). Galaxy Morphological Classification with Efficient Vision Transformer [arXiv:2110.01024 [astro-ph]]. https://doi.org/10.48550/arXiv.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., & Berg, J. v. d. (2008). Galaxy Zoo : Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey [arXiv:0804.4483 [astro-ph]]. *Monthly Notices of the Royal Astronomical Society*, *389*(3), 1179–1189. https://doi.org/10.1111/j.1365-2966.2008.13689.x

Mahabal, A., Sheth, K., Gieseke, F., Pai, A., Djorgovski, S. G., Drake, A. J., & Graham, M. J. (2017). Deep-learnt classification of light curves. *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*. https://doi.org/10.1109/ssci.2017.8280984

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Jia, Y., Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, … Xiaoqiang Zheng. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems [Software available from tensorflow.org]. https://www.tensorflow.org/

Martinazzo, A., Espadoto, M., & Hirata, N. (2020). Deep Learning for Astronomical Object Classification: A Case Study: *Proceedings of the*

15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, 87–95. https://doi.org/10.5220/0008939800870095

Misra, D., Mishra, S., & Appasani, B. (2018). Advanced Image Processing for Astronomical Images [arXiv:1812.09702 [cs]]. Retrieved March 1, 2023, from http://arxiv.org/abs/1812.09702

Thuan, T. X., & Gunn, J. (1976). A new four-color intermediate-band photometric system. *Publications of the Astronomical Society of the Pacific*, *88*(524), 543.

Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., & Yu, T. (2014). Scikit-image: Image processing in python. *PeerJ*, *2*, e453.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Walmsley, M., Lintott, C., Géron, T., Kruk, S., Krawczyk, C., Willett, K. W., Bamford, S., Kelvin, L. S., Fortson, L., Gal, Y., Keel, W., Masters, K. L., Mehta, V., Simmons, B. D., Smethurst, R., Smith, L., Baeten, E. M., & Macmillan, C. (2021). Galaxy Zoo DECaLS: Detailed visual morphology measurements from volunteers and deep learning for 314 000 galaxies. *Monthly Notices of the Royal Astronomical Society*, *509*(3), 3966–3988. https://doi.org/10.1093/mnras/stab2093

Walmsley, M., Lintott, C., Géron, T., Kruk, S., Krawczyk, C., Willett, K. W., Bamford, S., Kelvin, L. S., Fortson, L., Gal, Y., Keel, W., Masters, K. L., Mehta, V., Simmons, B. D., Smethurst, R., Smith, L., Baeten, E. M., & Macmillan, C. (2022). Galaxy Zoo DECaLS: Detailed visual morphology measurements from volunteers and deep learning for 314 000 galaxies., *509*(3), 3966–3988. https://doi.org/10.1093/mnras/stab2093

Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., & Schmidt, L. (2022). Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time.

Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., & Vajda, P. (2020). Visual transformers: Token-based image representation and processing for computer vision.

APPENDIX A

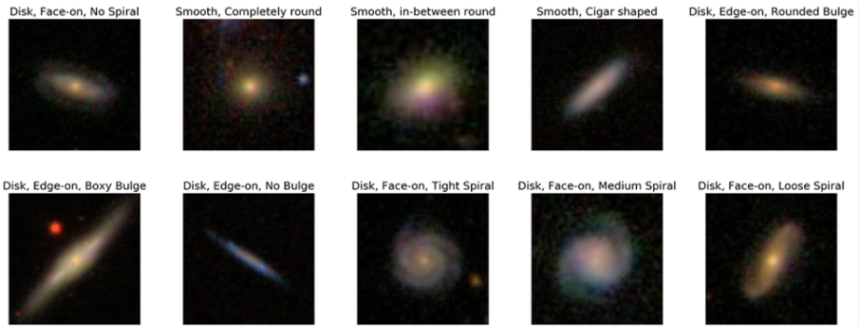

Figure 27: Example images from Galaxy10 SDSS dataset Retrieved from astroNN (Kollmeier et al., 2017; Lintott et al., 2008)

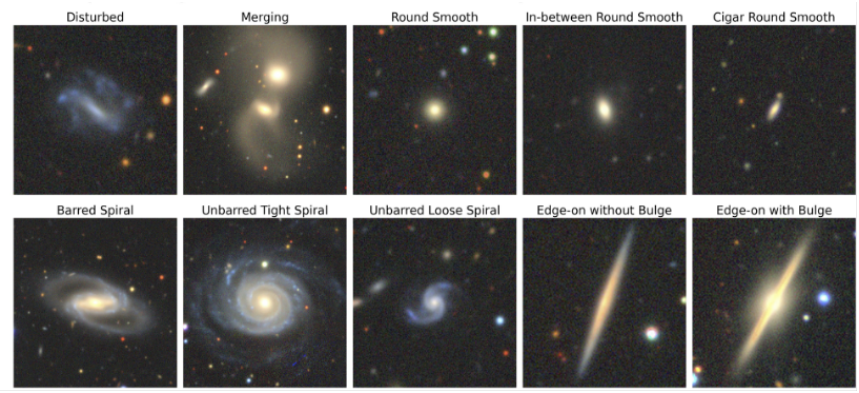

Figure 28: Example images from Galaxy10 DECals dataset Retrieved from astroNN(Dey et al., 2019; Walmsley et al., 2022)