Morphological Classification of Galaxies using SimCLR and Autoencoders

G036 (s1918275, s1957945, s1950841)

Abstract

With the rapid advancement in technology, telescopes are becoming more sophisticated and are able to gather vast amounts of data from galaxies. Unfortunately, for data to be fully useful, the processing of it must also be improved. In the past, classification of galaxies was conducted solely based on their appearance, by volunteers of citizen science, which was a time-consuming process. In response to this, scientists have started incorporating deep learning techniques such as Convolutional Neural Networks (CNNs) into their classification methods. However, a lot of previous research conducted performed galaxy classification in 2-way, 3-way, 4-way, 5-way or 10-way class, each using labelled dataset of different sizes. In this study, we propose 2 new models, Contrast Learning and Autoencoders, which are self-supervised learning models and will be using a much larger dataset of unlabelled data from Sloan Digital Sky Survey (SDSS). The models will then be fine tuned using the labelled dataset from Galaxy Zoo and, their accuracies will be compared to the past research to determine if using self supervised learning can outperform CNN on a 10-class labelled dataset.

1. Introduction

First introduced in 1926 by Edwin Hubble, morphological classification of galaxies refers to the categorisation of galaxies based on their visual appearance, structure and shape, as observed from Earth. As galaxies exhibit a wide variety of different structures, they are mainly grouped into 4 classes: spirals, barred spirals, elliptical and irregulars. This system provided a simple yet effective way of understanding the different types of galaxies and their characteristics.

The Hubble classification, also known as the Hubble Tuning Fork Diagram, is an important tool in modern astronomy and helps scientists and astronomers to understand the structure and evolution of the galaxies. It also allows them to study the relationship between the morphological features such as shape, size and luminosity of galaxies and the underlying processes that drive their formation. This information sheds light onto the formation of the stars, the universe as a whole and how they change over time and what elements are required for life, further advancing research in extraterrestrial life.

Throughout history, classification of galaxies was based solely on the visual appearance of the galaxies(Conselice, 2014). The Galaxy Zoo project, founded in July 2007, was created to allow volunteers to help quickly classify the galaxies in the Sloan Digital Sky Survey (SDSS) based on their shape. However, manually examining the images to classify the visual appearance of galaxies is a tedious and time consuming process. Although the number classifications performed by humans increased greatly, the actual classification itself was still slow. Additionally, it was slowly becoming impossible to perform classification due to the sheer volume of more detailed data produced by newer telescopes and much of the potentially useful data remained unused and unexplored (Barchi et al., 2020). In the past, the most common techniques used by astronomers were Fourier Transforms for temporal analysis, Least-Square Regressions or X^2 goodness of fit. However, even these traditionals methods were misused, leading to confusions when comparing studies (Feigelson & Babu, 2004). Hence, researchers developed automated classification techniques such as Convolutional Neural Networks (CNNs), Random Forest Classifiers and Ensemble Classifiers.

In (Cavanagh et al., 2021), the author developed both a 3 way(elliptical, lenticular, spiral) and 4 way(elliptical, lenticular, spiral, irregular) CNN architectures, achieving accuracies of 83% and 81% respectively. They also developed a 2 way/binary classification between all 4 classes and their model showed that it was easier to distinguish between ellipticals and spirals than spirals and irregulars. Next, they combined their binary classification model with their 3-way and 4-way classification in a hierarchical classification but performed poorly on the 4-way model, achieving an accuracy of only 65%, while their 3-way model achieved an accuracy of 81%. Finally, they applied their model on a small sample of the Galaxy Zoo images and achieved an accuracy of 92%, 82%, 77% in binary, 3 way and 4 way classifications respectively.

The authors in (Cheng et al., 2020a) compared several supervised machine learning techniques; CNNs, Random Forest, Logistic Regression and Support Vector Machines (SVMs) for classifying galaxies from the Dark Energy Survey (DES) dataset, using visual classification of 2800 galaxies from Galaxy Zoo 1. Out of all the methods, the authors found that CNNs performed the best, with an accuracy of 0.99 on a 2 class classification (ellipticals and spirals).

In (Gharat & Dandawate, 2022), the authors proposed using CCNs, with each convolution kernel extracting multiple features, which are then injected into the classifier to obtain

10 galaxy classes, achieving an accuracy of 84.73%. This paper is used as out baseline

As most of the research is performed in classes of 2-way, 3-way, 4-way, 5-way or 10-way, using supervised learning, it is difficult to perform some comparisons in between them. Moreover, each of the experiments are performed on a different size of dataset. In this paper, we propose two different models and perform several investigations to determine which model can outperform the baseline.

In the first model, we propose to use the self-supervising learning methods; SimCLR Contrast Learning, while in the second model, we will be employing Autoencoders. For both models, we will be using the unlabelled dataset from Sloan Digital Sky Survey (SDSS). We will further fine tune our model by using the 10 class labelled dataset from Galaxy Zoo.

The goal of these two experiments is to determine if self-surpervised learning can provide a higher accuracy than supervised learning, and which of the self-supervised learning method would be best suited.

2. Data set and task

The dataset we used in this paper is a combined dataset, where the image data came from Sloan Digital Sky Survey (SDSS) and the data labels came from the Galaxy Zoo project, as well as Galaxy 10 which is a rigorously filtered subset of Galaxy Zoo project.

2.1. SDSS Dataset

Sloan Digital Sky Survey (SDSS) is a survey aimed to produce images (indicate shape) and spectrums (indicate distance) of the galaxies over a large area of the sky by the 2.5m f/5 telescope from the Apache Point Observatory, New Mexico (York et al., 2000). The survey starts from the year 2000, until 2012, the survey had 9 data releases which covers over 35% of the sky(Anderson et al., 2012). In this paper, we're using a subset of the SDSS dataset picked by Galaxy Zoo, which contains 243,434 images of the brightest galaxies in the SDSS dataset.

2.2. Galaxy Zoo Project

The Galaxy Zoo project is a citizen science project which provides the morphological classification of galaxy images from the SDSS dataset(Willett et al., 2013). The project largely speeds up the morphological classification of the huge amount of existing galaxy images from SDSS. However, the public volunteering characteristic also makes the labels for each galaxy highly messy and unreliable, which makes the deep learning models hard to achieve a promising result on the labels provided by Galaxy Zoo.

2.3. Galaxy 10 SDSS Dataset

Galaxy 10 SDSS dataset is a subset of the galaxy zoo project. The dataset only selects the galaxy images with more than 55% of the votes agreeing on the same classes(Ast). The dataset contains 21758 coloured galaxy images which have a size of 69 * 69 pixels. Examples of the data are shown in figure 1. The class distribution is shown in table 1.

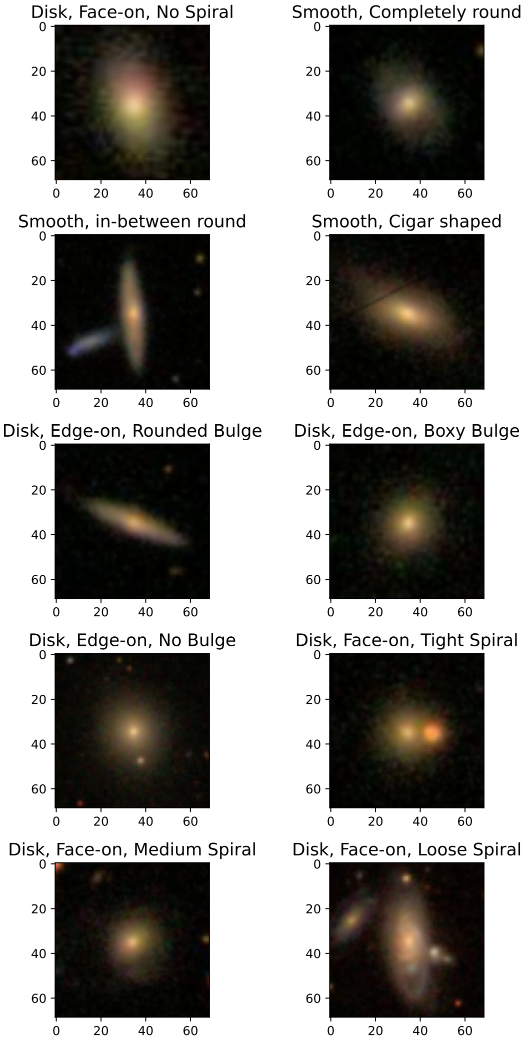

Disk, Face-on, No Spiral
Smooth, Completely round
Smooth, in-between round
Smooth, Cigar shaped
Disk, Edge-on, Rounded Bulge
Disk, Edge-on, Boxy Bulge
Disk, Edge-on, No Bulge
Disk, Face-on, Tight Spiral
Disk, Face-on, Medium Spiral
Disk, Face-on, Loose Spiral
0
20
40
60
0
20
40
60

Figure 1. Images of each Galaxy Class

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

Table 1. Galaxy 10 Dataset Class Distribution

2.4. Data Split

In this paper, we will use the 21758 rigorous filtered labelled data from the Galaxy 10 SDSS dataset as the labelled dataset, and 243,434 image data from the SDSS selected by the galaxy zoo as the unlabeled dataset.

The labelled dataset will be split into training, validation and testing datasets in the ratio 70:15:15.

2.5. Task

In this paper, we aim to develop an efficient deep learning-based approach to classify the galaxies' images according to their morphological features in a self-supervised manner. We aim to outperform the accuracy of the model proposed in (Gharat & Dandawate, 2022) with the same amount of labelled data used.

3. Methodology

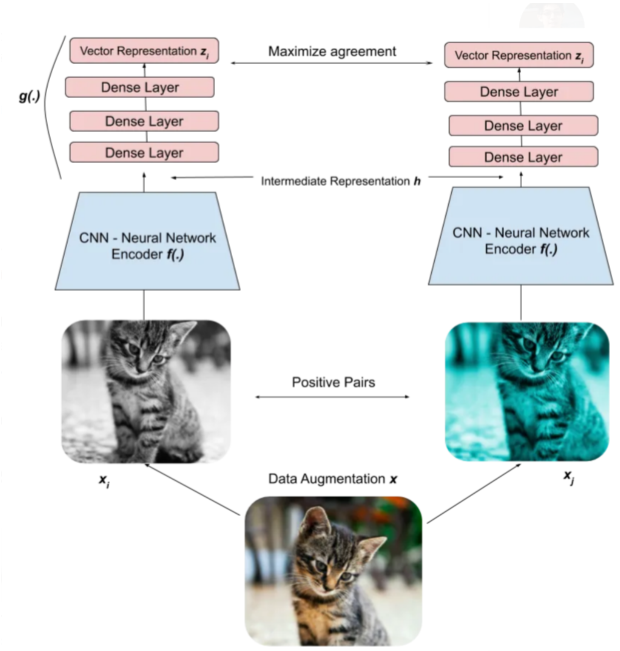
3.1. Convolutional Neural Network

Convolutional Neural Network (CNN) is a special type of Artificial Neural Network (ANN). Compared with normal ANNs, CNNs will have a special type of layers called "convolutional layers". At a high level, convolutional layers will extract some features from the input data. As more and more convolutional layers have been passed, the features that are found by the convolutional layers will be more specific and higher level. Taking an image of a dog as an example, the very first convolutional layers in the network could only extract low-level features like horizontal and vertical edges. As more convolutional layers have been passed, higher-level features like tails, and mouths will be extracted from the data. Finally, the Neural Network can perform classification or detection tasks based on the features extracted from convolutional layers.

CNN is widely used in image-processing tasks mainly because the convolutional layers in the CNN could reduce high-dimension images to extract features of the image while not losing any information(Sharma et al., 2018). Compared with normal ANN, when dealing with high-dimension data will result in a huge amount of trainable parameters.

3.2. Self-supervised Learning

Self-supervised learning is a machine learning technique which makes the model learn to represent the data in a meaningful way by solving a few pretext tasks with unlabeled data. Self-supervised learning is particularly useful when there is a limited amount of labelled data with a huge amount of unlabelled data, where pre-train the network with a self-supervised learning manner then fine-tuning the network with limited supervision could create a similar performance compared with the model trained with fully supervised that uses 10 times more labelled data (Chen et al., 2020c). In this paper, we will investigate the performance of two self-supervised learning architectures, namely SimCLR and autoencoder, in the task of galaxy morphological classification.



g(.)
Vector Representation z_i
Maximize agreement
Vector Representation z_j
Dense Layer
Dense Layer
Dense Layer
Dense Layer
Dense Layer
Dense Layer
Intermediate Representation h
CNN - Neural Network Encoder $f(.)$
CNN - Neural Network Encoder $f(.)$
Positive Pairs
x_i
Data Augmentation x
x_j

Figure 2. An overview of the SimCLR framework, photo by (Sim)

3.2.1. SimCLR

SimCLR is a network designed for contrast learning of visual representations(Chen et al., 2020b). Contrast learning is a type of self-supervised learning, where the network learns to produce meaningful representations by maximizing the agreement between positive picture pairs of pictures. The positive pair of images are generated by different and heavy augmentations from the same image. As shown in figure 2 below, the picture of the cat is augmented differently to form different views of the same image, the two different views are considered positive pairs. The positive pair of images are then fed to the network, and the network will be trained to minimise the difference in the representations of different views of the same image. (Chen et al., 2020b) shows by using the same amount of labelled data, the SimCLR network outperforms the supervised counterpart in 9 among 12 datasets.

3.2.2. Autoencoder

Autoencoder is another self-supervised representation learner for computer vision. The autoencoder learns to produce meaningful representation by maximising the agreement between the original input image and the image reconstructed from the compressed representation that came out of the encoder (He et al., 2022). (Cheng et al., 2020b) applied autoencoders to the task of detecting a strong gravitational lens effect, the autoencoder network successfully outperform the CNN baseline, showing the autoencoder is able to deal with classification tasks with small intra-class differences.

To use the autoencoder network for classification, the decoder will be replaced with a dense layer. Then the entire network (the encoder and the dense layer) will fine-tune the limited amount of labelled data.

4. Experiments

The model proposed by (Gharat & Dandawate, 2022) uses the well-established VGG8 architecture which is comprised of interweaved convolution blocks and MaxPooling layers. This architecture has been commonly used to solve various image classification tasks and thus provides a solid baseline for this study.

Following the original paper's setup, the dataset was imported using a library called astroNN which contains 21785 images at a resolution of 69x69. Then the images were normalized by dividing each pixel value by 255 which scales each pixel to a range between 0-1. The data is then split using a 70:15:15 ratio for train, testing and validation data respectively. The Keras framework was used to construct the model, optimizing the categorical cross-entropy loss with the Adam optimizer. This model was trained for 20 epochs with the rest of the fields being kept as default.

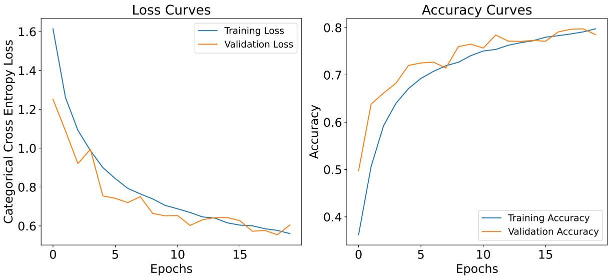
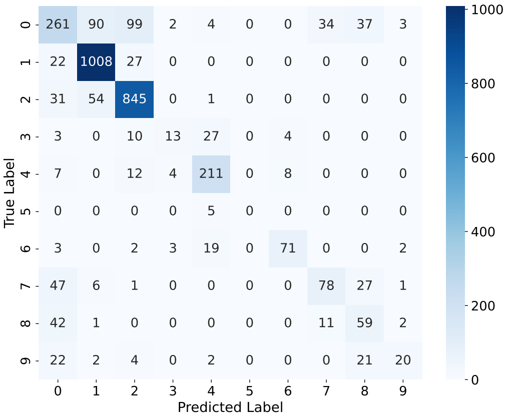


Figure 3. Accuracy and Loss Curves During Training



	0	1	2	3	4	5	6	7	8	9
0	261	90	99	2	4	0	0	34	37	3
1	22	1008	27	0	0	0	0	0	0	0
2	31	54	845	0	1	0	0	0	0	0
3	3	0	10	13	27	0	4	0	0	0
4	7	0	12	4	211	0	8	0	0	0
5	0	0	0	0	5	0	0	0	0	0
6	3	0	2	3	19	0	71	0	0	2
7	47	6	1	0	0	0	0	78	27	1
8	42	1	0	0	0	0	0	11	59	2
9	22	2	4	0	2	0	0	0	21	20

Figure 4. Confusion Matrix

The architecture proposed by (Gharat & Dandawate, 2022) is hindered by class imbalance in the dataset, which results in limited data available for certain classes. For instance, Class 5 (Disk, Edge-on, Boxy Bulge) only has 17 images, which are further divided into training, testing, and validation data. This lack of data is reflected in the results of the baseline model, as shown in Figure 4. The model was unable to make any predictions for Class 5 due to the limited data available, which was caused by the filtering process used in astroNN. Public voting was used to obtain the labels for these images, but only those on which 55% of the voting population agreed were considered in the dataset. To overcome this limitation, we propose a self-supervised architecture using SimCLR contrastive learning and/or Autoencoders to improve classification accuracy by including more data from GalaxyZoo, particularly the lower confidence data.

While convolutional neural networks (CNNs) have been a dominant approach in computer vision in the past, recent studies have demonstrated the superiority of Vision Transformers (ViTs) in image classification tasks (Borhani et al., 2022)(Chen et al., 2021)(Dosovitskiy et al., 2020)(Zhao et al., 2021)(Paul & Chen, 2021). Our proposal seeks to combine the strengths of both ViTs and CNNs by incorporating the ViT architecture into the feature extraction module, alongside the proven CNN design. This approach is expected to yield higher accuracy and improved model performance, while still keeping the number of parameters relatively low.

5. Interim conclusions

The combination of ViTs and CNNs has been proven to enhance image classification performance, according to (Chen et al., 2021). Furthermore, SimCLR contrastive learning has demonstrated its potential to reach or surpass state-of-the-art results through semi-supervised methods (Chen et al., 2020a). The study by (Gharat & Dandawate, 2022) is unique in that it is the only one to address the galaxy classification task with 10 classes. Previously, this task was tackled with a lower number of classes, 7 in (Kalvankar et al., 2020) and 5 in (Zhu et al., 2019). Shreyas et al. revealed that the model presented by Zhu et al. experienced a drop in accuracy from 95.2% to 57.1% with the addition of two new classes. Given the increase of classes to 10, it is expected that similar decreases in performance will happen, making it necessary to conduct further research with SimCLR or autoencoders. The baseline model achieved an accuracy of 78.5% using the VGG8 model architecture, which is similar to the results reported by (Gharat & Dandawate, 2022). However, (Kalvankar et al., 2020) reported an accuracy of 75.5% with a VGG16 model and fewer classes, which contradicts what would be expected with fewer class labels. The poor performance might be due to the skewed data, as the classes with better baseline results contain more samples (Table 1 and Figure 4).

6. Plan

As we stated in section 3.4 above we aim to develop a deep-learning model for galaxy morphological classification with self-supervised pretraining, the model could outperform the baseline model in Galaxy 10 dataset with the same amount of labelled data used. The project has a few milestones shown below, we also list a few extra milestones we will work on if we finish the previous milestones in time, which also act as backup plans.

1. Choose the encoder architecture of SimCLR and autoencoder.
2. Train the self-supervised models on Galaxy 10 dataset.
3. Train the self-supervised models on unlabeled SDSS dataset with Galaxy 10 dataset.

Extra Milestones/ Backup Plans

1. Add vision transformers in architectures of encoders for both the baseline model and self-supervised models.
2. Try the same work on Galaxy10 DECals Dataset, which is another galaxy morphological classification dataset. But it has a higher resolution and more balanced class data distribution compared with the Galaxy 10 SDSS dataset.

References

Galaxy10 SDSS Dataset. https://astronn.readthedocs.io/en/latest/galaxy10sdss.html, accessed 26 January 2023.

SSL explain. https://towardsdatascience.com/understanding-contrastive-learning-d5b19fd96607, accessed 31 January 2023.

Anderson, Lauren, Aubourg, Eric, Bailey, Stephen, Bizyaev, Dmitry, Blanton, Michael, Bolton, Adam S, Brinkmann, Jon, Brownstein, Joel R, Burden, Angela, Cuesta, Antonio J, et al. The clustering of galaxies in the sdss-iii baryon oscillation spectroscopic survey: baryon acoustic oscillations in the data release 9 spectroscopic galaxy sample. *Monthly Notices of the Royal Astronomical Society*, 427(4):3435–3467, 2012.

Barchi, P.H., de Carvalho, R.R., Rosa, R.R., Sautter, R.A., Soares-Santos, M., Marques, B.A.D., Clua, E., Gonçalves, T.S., de Sá-Freitas, C., and Moura, T.C. Machine and deep learning applied to galaxy morphology - a comparative study. *Astronomy and Computing*, 30:100334, 2020. ISSN 2213-1337. doi: https://doi.org/10.1016/j.ascom.2019.100334. URL https://www.sciencedirect.com/science/article/pii/S2213133719300757.

Borhani, Yasamin, Khoramdel, Javad, and Najafi, Esmaeil. A deep learning based approach for automated plant disease classification using vision transformer. *Sci. Rep.*, 12(1):11554, July 2022.

Cavanagh, Mitchell K, Bekki, Kenji, and Groves, Brent A. 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, 2021.

Chen, Chun-Fu, Fan, Quanfu, and Panda, Rameswar. Crossvit: Cross-attention multi-scale vision transformer for image classification, 2021. URL https://arxiv.org/abs/2103.14899.

Chen, Ting, Kornblith, Simon, Norouzi, Mohammad, and Hinton, Geoffrey. A simple framework for contrastive learning of visual representations, 2020a. URL https://arxiv.org/abs/2002.05709.

Chen, Ting, Kornblith, Simon, Norouzi, Mohammad, and Hinton, Geoffrey. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020b.

Chen, Ting, Kornblith, Simon, Swersky, Kevin, Norouzi, Mohammad, and Hinton, Geoffrey E. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020c.

Cheng, Ting-Yun, Conselice, Christopher J, Aragón-Salamanca, Alfonso, Li, Nan, Bluck, Asa F L, Hartley, Will G, Annis, James, Brooks, David, Doel, Peter, García-Bellido, Juan, James, David J, Kuehn, Kyler, Kuropatkin, Nikolay, Smith, Mathew, Sobreira, Flavia, and Tarle, Gregory. Optimizing automatic morphological classification of galaxies with machine learning and deep learning using Dark Energy Survey imaging. *Monthly Notices of the Royal Astronomical Society*, 493(3):4209–4228, 02 2020a. ISSN 0035-8711. doi: 10.1093/mnras/staa501. URL https://doi.org/10.1093/mnras/staa501.

Cheng, Ting-Yun, Li, Nan, Conselice, Christopher J, Aragón-Salamanca, Alfonso, Dye, Simon, and Metcalf, Robert B. Identifying strong lenses with unsupervised machine learning using convolutional autoencoder. *Monthly Notices of the Royal Astronomical Society*, 494 (3):3750–3765, 2020b.

Conselice, Christopher J. The evolution of galaxy structure over cosmic time. *Annual Review of Astronomy and Astrophysics*, 52(1):291–337, 2014. doi: 10.1146/annurev-astro-081913-040037. URL https://doi.org/10.1146/annurev-astro-081913-040037.

Dosovitskiy, Alexey, Beyer, Lucas, Kolesnikov, Alexander, Weissenborn, Dirk, Zhai, Xiaohua, Unterthiner, Thomas, Dehghani, Mostafa, Minderer, Matthias, Heigold, Georg, Gelly, Sylvain, Uszkoreit, Jakob, and Houlsby, Neil. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. URL https://arxiv.org/abs/2010.11929.

Feigelson, Eric D and Babu, GJ. Statistical challenges in modern astronomy. *arXiv preprint astro-ph/0401404*, 2004.

Gharat, Sarvesh and Dandawate, Yogesh. Galaxy classification: a deep learning approach for classifying sloan digital sky survey images. *Monthly Notices of the Royal Astronomical Society*, 511(4):5120–5124, 2022.

He, Kaiming, Chen, Xinlei, Xie, Saining, Li, Yanghao, Dollár, Piotr, and Girshick, Ross. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Kalvankar, Shreyas, Pandit, Hrushikesh, and Parwate, Pranav. Galaxy morphology classification using efficientnet architectures, 2020. URL https://arxiv.org/abs/2008.13611.

Paul, Sayak and Chen, Pin-Yu. Vision transformers are robust learners, 2021. URL https://arxiv.org/abs/2105.07581.

Sharma, Neha, Jain, Vibhor, and Mishra, Anju. An analysis of convolutional neural networks for image classification. *Procedia computer science*, 132:377–384, 2018.

Willett, Kyle W, Lintott, Chris J, Bamford, Steven P, Masters, Karen L, Simmons, Brooke D, Casteels, Kevin RV, Edmondson, Edward M, Fortson, Lucy F, Kaviraj, Sugata, Keel, William C, et al. Galaxy zoo 2: detailed morphological classifications for 304 122 galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 435(4):2835–2860, 2013.

York, Donald G, Adelman, J, Anderson Jr, John E, Anderson, Scott F, Annis, James, Bahcall, Neta A, Bakken, JA, Barkhouser, Robert, Bastian, Steven, Berman, Eileen, et al. The sloan digital sky survey: Technical summary. *The Astronomical Journal*, 120(3):1579, 2000.

Zhao, Yucheng, Wang, Guangting, Tang, Chuanxin, Luo, Chong, Zeng, Wenjun, and Zha, Zheng-Jun. A battle of network structures: An empirical study of cnn, transformer, and mlp, 2021. URL https://arxiv.org/abs/2108.13002.

Zhu, Xiao-Pan, Dai, Jia-Ming, Bian, Chun-Jiang, Chen, Yu, Chen, Shi, and Hu, Chen. Galaxy morphology classification with deep convolutional neural networks. *Astrophysics and Space Science*, 364(4), apr 2019. doi: 10.1007/s10509-019-3540-1. URL https://doi.org/10.1007%2Fs10509-019-3540-1.