Astronomy & Astrophysics manuscript no. output ©ESO 2022
April 19, 2022

Galaxy Morphological Classifications Using Machine Learning

Sithila Premnath[1]
Supervised by
Dr. Laura Parker[2]

[1] McMaster University
e-mail: premnats@mcmaster.ca
[2] McMaster University
e-mail: lparker@mcmaster.ca

April 2022

ABSTRACT

This review explores the current state of machine learning (ML), data analysis and neural networks in galaxy morphological classifications. The application of ML models has become an increasingly popular tool that allows for reliable galaxy classification and merger identification. Observational multi-wavelength data taken from 'Big-Data' surveys (Sloan Digital Sky Survey (SDSS), Dark Energy Survey (DES), etc.) has been used to train the two machine learning models; Supervised (SML) and Unsupervised (UML). In particular, Convolutional Neural Networks (CNNs), a type of supervised deep learning are a common tool used in research for galaxy classification. However, the computational costs of SML deters the handling of large datasets. Statistical machine learning methods, shows an average 95% accuracy (across the literature) - when compared to Galaxy Zoo - for Spiral and Elliptical galaxy classification but due to visual inconsistencies in mergers and the remoteness of stars, there is an inaccuracy in their classifications. Unsupervised methods are increasing in use due to their ability to deduce a network of similarities and produce classification rules that are largely in agreement with manual analysis. Moving forward, a mix of UML and SML may advance how ML algorithms and morphological research is conducted, but computational costs will hinder any exponential progress.

Key words. machine learning – galaxy morphology– neural networks – deep learning

1. Introduction

Galaxy morphology is the study of the shape, structure and the physical processes galaxies undergo and is a fundamental part of observational cosmology. The term morphology refers to the observed shapes of galaxies and the first attempt to categorize these morphology's was made by Edwin Hubble's "Tuning Fork" Hubble (1936). Hubble proposed a system of classifications (Elliptical (E), Spiral (S), lenticular or irregular (Ir)). This was thought to be a chronological sequence with bulge-dominated galaxies (E) at early epochs of time also known as Early-Type Galaxies (ETG). Galaxies with prominent disk components (S) were called Late-Type Galaxies (LTG). Hubble further classified Spirals into barred, Spirals with a central bar shaped structure composed of stars and unbarred (no central bar). The inclusion of T-Type de Vaucouleurs (1963) equipped astronomers with the ability to classify galaxies that do not fall under the three main groups. The T-type is a number system that starts at -6 (Ellipticals) and ends at +10 (Spirals and irregulars).

With large advancements in technology the ease with which we observe galaxies is considerably high. However, astronomy is an extremely data rich field and with new telescopes and instruments on board of satellites (e.g. the Hubble Space Telescope (HST)) we obtain massive datasets Barchi et al. (2020) that are impossible to quantify through human visual inspection. Another limitations of human classification comes from the transition states between T-Types de Vaucouleurs (1963). The inclusion of T-Type equipped astronomers with the ability to classify galaxies that do not fall under the three main groups. The T-type is a number system that starts at -6 (Ellipticals) and ends at +10 (Spirals and irregulars).

The literature reviewed in this paper analyzes galaxy data of nearby regions of space - low redshifts ($z < 1$), as structural information is lost at $z > 1$ and T-Types are harder to observe. Efforts have been made to tackle the data problem in the form of Galaxy Zoo (GZ; C. Lintott et al. (2017), a citizen science project in which visual classification is implemented to distinguish between Spirals and Ellipticals. Unfortunately, the issue

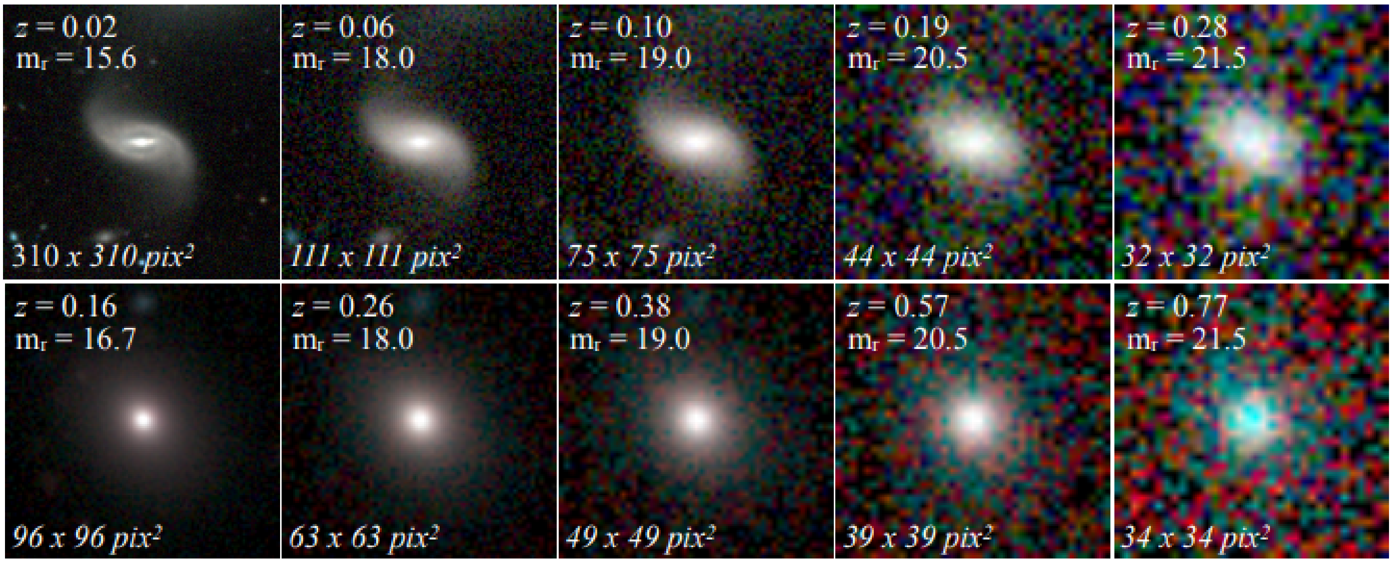


Fig. 1: Cutouts for Spiral (top) and Elliptical (bottom) with increasing redshift and apparent magnitude. Physical properties such as Spiral tail and Elliptical bulge are lost at high redshift, these images are used as the training set in this CNN J. Vega-Ferrero et al. (2021).

with open to public (amateur) classifications is that galaxies tend to be incorrectly labelled. The proportion of Elliptical galaxies mislabelled in comparison to Spirals makes implementing GZ data in training machines a difficult task H. Domínguez Sánchez (2018). The availability of various T-Type options (barred, irregular, bulge dominance, roundness, etc.) lead to many misclassifications, especially for unusual looking galaxies (e.g. Irregular, Lenticular, etc.). Even though the efforts of GZ have enhanced the classification procedure, the size of the various Big-Data surveys impedes thorough systematization.

To combat this limitation, machine learning and deep learning (DL) techniques can pave the way towards a more robust and reliable classification system. Machines have two approaches to learning information, one being the supervised method where algorithms are trained by pre-labelled datasets and human intervention to produce classifications. In particular, the use of Convolutional Neural Networks (CNNs) in morphological research, is frequently implemented due to its ability to classify raw galaxy images without the need to process and clean data (pre-process) Barchi et al. (2020). CNNs have surpassed many supervised automated methods, such as logistic regression, random forests, decisions trees. in regard to accurate measurements and computational costs. The drawback with most supervised methodology is that it requires a considerable amount of pre-labelled data to train algorithms J. Vega-Ferrero et al. (2021). Additionally, morphological classifications rely on the ability to differentiate between ETGs and LTGs and with large samples from big-data surveys, there is difficulty in the sorting process Barchi et al. (2020). Fortunately, the access to various surveys allows for transfer learning H. Domínguez Sánchez (2018), which is a learning technique that adapts previous knowledge onto a new problem set. On the other hand, UML algorithms are the optimal approach in Big-Data morphological analysis as there is no need to train algorithms with visually labelled datasets. In principle, UML has the ability to rapidly and autonomously create classification schemes by compressing galaxy imaging into clusters called, "morphological clusters" G. Martin et al. (2019). These groupings are then benchmarked against current visual classifications (SDSS) and can be used to train SML algorithm.

In CNN image recognition, images are first broken down into sections, light or dark areas are grouped separately, whereby complex structural information is categorized by each node and finally the network outputs a classification. Each node is given a weight that is manipulated in accordance with developments in the model. The weight is a link between attributes in the input image and represents the strength of its relationship with the final classification. Deeper into the network, nodes have a higher weighting. Alternatively, the ability to parse through massive amounts of data without the need for labelled data gives unsupervised learning an edge above the usual supervised approach.

This paper focuses on how machine learning and neural networks are becoming the standard tool in galaxy observation. Specifically, I will go over the limitations of both supervised and unsupervised methods and how a collaboration of the two will advance morphological classifications. This document is organized as follows: Section 2 describes the datasets used. Section 3 describes the learning process of the supervised and unsupervised methods. Section 4 describes how well the algorithms performed and discusses the accuracy's between traditional machine learning (TML) and DL methodology's. I evaluate and analyze the possible applications of each research approach and summarize their key finding's in section 5.

2. Data Sets

This literature reviewed in this paper relies on data taken from three big data surveys: The Dark Energy Survey (DES), the Sloan Digital Sky Survey (SDSS) and the Galaxy Zoo.

2.1. Dark Energy Survey

The DES is a wide-field optical imaging survey that covers approximately 1/8 of the sky E. L. Neilsen et al. (2019) and is equipped with the Dark Energy Camera Ting-Yun Cheng (2019) that is sensitive to a highly redshifted universe (~ 650 to ~ 900 nm in the red band). Its ability to observe red wavelengths gives us high quality images of distant galaxies compared to previous surveys. At ½ square degree the camera captures a coadd (tile) of size 10000x10000 pixels and each coadd is used to train the neural networks evaluated in this review.

Fig. 2: Deep field image 226 million galaxies surveyed by the DES

Low redshifts are paramount when training machine learning algorithms as galaxy structure is attenuated at high redshifts. CNNs trained on large galaxy catalogues at apparent magnitude brighter than 17.7 would fail to classify considerably fainter galaxies. One remedy is to visually classify fainter galaxies; however, this is a task that is subject to bias and lack of resources (as seen in GZ). A simpler alternative is to simulate what actual DES galaxies would look like given higher redshifts J. Vega-Ferrero et al. (2021).

In figure 1, original cutouts are taken at their original redshifts and noise is added onto each simulated galaxy. The apparent magnitude increases up to an m_r of 21.5 to ensure the CNN is learning galaxies that are bright enough for the DES's detection threshold. In both cases of the LTG (z = 0.02) and ETG (z = 0.16) there are no distinguishable features by eye at large simulated apparent magnitudes. This process is carried out for each galaxy used in training this CNN.

2.2. Sloan Digital Sky Survey

The Sloan Digital Sky Survey (SDSS) uses a five band (u-g-r-i-z) optical and near-infrared spectroscopic in addition to applications of redshift analysis. A 2.5m telescope measures light from distant galaxies and a 0.4m telescope observes relatively bright stars in our galaxy. The dataset used to train the algorithm Reza (2021) consists of 100,173 galaxy samples: 60,501 Spirals, 39,264 Ellipticals, 292 mergers and 116 stars.

2.3. Galaxy Zoo 2

GZ2 is a catalogue that contains ~ 240,000 galaxies taken from SDSS DR7 Legacy Survey with classifications conducted by citizen science volunteers. Labelling is done through a decision tree interface where volunteers answer questions for individual galaxy images [J. Vega-Ferrero et al. (2021)]. The GZ2 decision tree has 11 classification tasks with 37 possible responses (each question consists of two to seven possible answers). GZ2 contains a total number count of the votes, including a fractional representation of each answer. The final fractional votes for each class are the weighted mean of each individual vote; a more consistent answer gets a higher weighting Reza (2021). GZ2 elaborates a more sophisticated morphological process that includes features such as bars, bulges and Spiral arms.

3. Machine and Deep Learning

In this section, I will focus on three classification methods: clustering, CNNs and traditional SML. With big-data surveys, the accuracy and effectiveness of manual inspection becomes increasingly difficult. The implementation of machine learning algorithms can be traced back to the late 20th century O. Lahav et al. (1995) but even then the computational demand due to terabytes of data creates room for unreliability. Despite this, approaches through SML such as decision trees (DT), random forests (RF), artificial neural networks (ANN), principal component analysis (PCA), extra trees (ET) and K- nearest neighbours (KNN) boast an average Spiral and Elliptical classification of 95% but falls short in classifications of mergers and stars, ~ 40 % Reza (2021).

3.1. Supervised Learning

3.1.1. Traditional Supervised Learning

Non-parametric galaxy morphology is a computing approach that determines Concentration (C), Asymmetry (A), Smoothness (S), Entropy (H), and Gradient Pattern Analysis (GPA) metrics. By using a python based image processing and feature extraction algorithm called CyMorph, analysis is done through three major steps: producing a galaxy stamp, removing secondary objects and generating segmented images. Concentration is measured as $C = \log_{10}(R_1/R_2)$ where R_1 and R_2 are the outer and inner radii of a galaxy image around the centre. Asymmetry is the correlation between the original and rotated image. Smoothness defines the correlation between the original image and flux

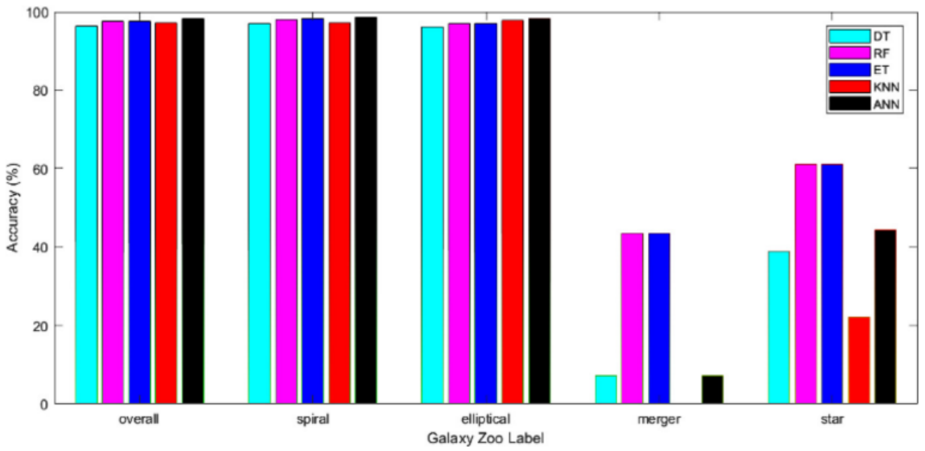


Fig. 3: Bar chart displaying classification accuracy's for different supervised methods Reza (2021)

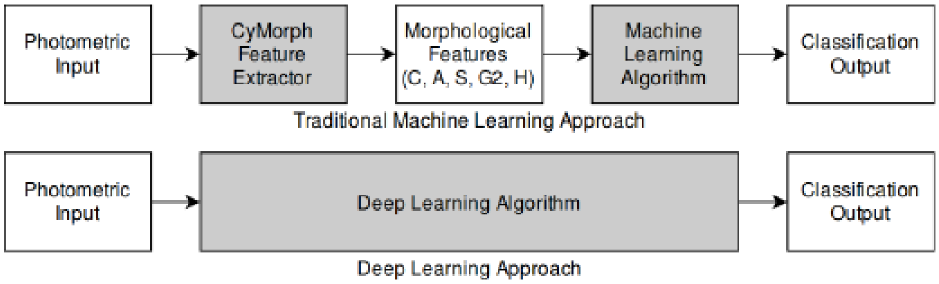

Photometric Input
CyMorph Feature Extractor
Morphological Features (C, A, S, G2, H)
Machine Learning Algorithm
Classification Output
Traditional Machine Learning Approach
Photometric Input
Deep Learning Algorithm
Classification Output
Deep Learning Approach

Fig. 4: Flow chart of a deep learning algorithm opposite a traditional machine learning approach. Barchi et al. (2020)

dispersion of the image which is how the gradients vary over the entire image. The GPA establishes a method to estimate local gradient properties in a given image and the entropy measures the distribution of pixel values in the image Barchi et al. (2020). CyMorph uses these parameters to create distinctions between morphological types (Fig. 4).

3.1.2. Convolutional Neural Networks

Convolutional Neural Networks are a type of deep learning that use convolutional layers (layers of neural nodes) to extract information as weights to form a classification set. The main advantage of CNNs is that they do not require a pre-processing procedure typical to traditional machine learning. The CNN I will explore in this paper uses fewer convolutional layers in addition to two types of pre-processed images that undergo two data preparation stages: stamp creation and image processing. Stamp creation takes the original coadd images from a size of 10000x10000 pixels to millions of 50x50 pixel 'postage stamp' images and rotates them by different angles which improves the CNN due to an increase in the sample of training images. This training set is then divided into two groups by using gradient analysis, a pattern recognition technique that differentiates object appearance and shape within an image Ting-Yun Cheng (2019). The second input takes the original images and runs it on a logarithmic scale, called log images. The architecture of the CNN in figure 4 starts with an input of a 50x50x3 dimension portion of an image which is filtered by three convolutional layers of sizes 3,3,2 Ting-Yun Cheng (2019) and channel sizes of 32, 64 and 128. A max pooling is applied to each convolutional layer and then a dropout of 0.5 is applied to combat data overfitting. Finally, the probabilities of a galaxy being either a Spiral or an Elliptical is given as the output. This particular CNN is accurate to 99% Ting-Yun Cheng (2019) and would be 100% if not for mismatches in Galaxy Zoo classifications (Spirals classified as Ellipticals). The results obtained by this paper, for the time being, contain the largest morphological classifications available for analysis Ting-Yun Cheng (2019).

3.2. Unsupervised Learning

Galaxy classification using unsupervised machine learning relies on clustering unseen graphical information from large graphical data into small groups of visually similar objects. With an abundance of data from surveys, the ability to adapt to rapidly evolving data and provide classifications is valuable due to the challenges faced in training algorithms. The technique employed has been shown to separate Elliptical and Spiral galaxies efficiently when compared to current morphological classification schemes

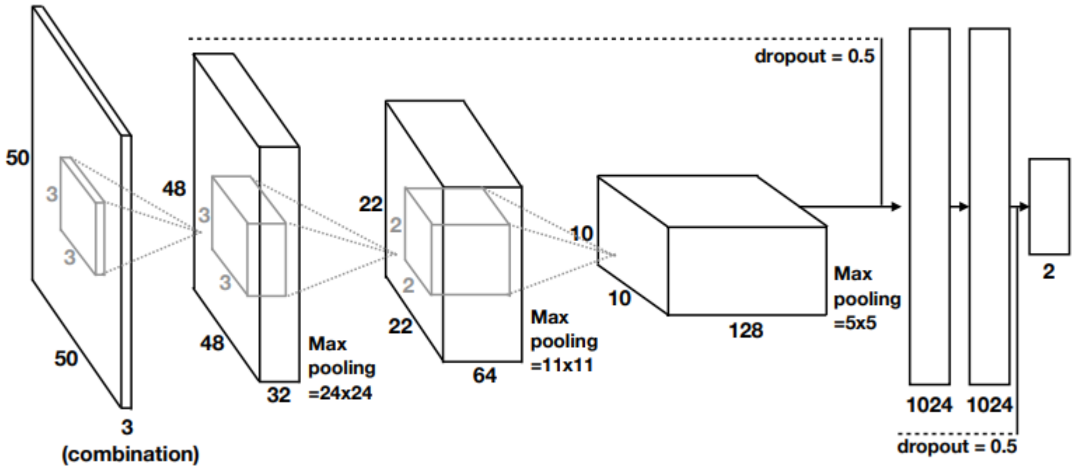


Fig. 5: Process in which a CNN algorithm predicts final outputs for Spiral and Elliptical galaxies Barchi et al. (2020)

Hocking et al. (2017). The main components of the algorithm are termed feature selection and feature extraction.

3.2.1. Clustering

Feature selection aims to automatically identify specific qualities in galaxies from pixel data. Redundancies such as galaxy rotation and scale from observation are removed to keep galaxies invariant and simple when training the algorithm. Feature extraction is depicted in figure 5 where the UML extracts features of galaxies and clusters information into feature vectors to produce a library of patch types. A growing neural gas (GNG; B. Fritzke et al. (1995)) algorithm is applied, which is a process that finds an optimal representation of the feature vectors and then creates a topological mapping. This mapping is run through hierarchical clustering (HC) that works bottom up to reduce the feature vectors into object feature vectors that represent the frequency of different patch types. Once these vectors are cleaned to exclude objects less than 15 pixels (irregularities that can be disregarded), k-means clustering is implemented to separate object vectors into k morphological clusters. Expert visual classifications are cross-matched with each cluster into one of three broad Hubble morphological types G. Martin et al. (2019): Elliptical galaxies, S0/Sa galaxies and Spiral galaxies.

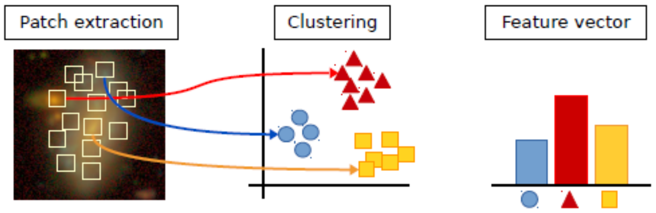


Fig. 6: Clustering methods (UML), divides patch data by group features and produces a library of distinct patch types (feature vector) Reza (2021)

4. Morphological Cataloguing Accuracy

Methods to analyze clustered data can be done by measuring the star formation rate (SFR) and stellar mass distribution as a function of morphological type. Figure 7 shows the evolution of galaxy stellar mass functions (left column) and the evolution of morphological fractions as a functions of four redshift bins (right).The lightly dotted lines represent Spiral,Elliptical and S0/Sa curves from previous studies. The notable differences between the solid and the dotted curves is due to differing definitions of morphological typing G. Martin et al. (2019). Despite this, the general trend in Spiral, Elliptical and S0/Sa galaxy fractions are constant. Even though S0/Sa and Ellipticals share similar mass functions at low redshifts, Ellipticals dominate in stellar mass at the early universe indicating there is a distinct evolutionary channel that forms S0/Sa galaxies G. Martin et al. (2019).

Spiral galaxies have stars that inhabit a well defined main sequence, while Ellipticals contain stars that dominate the cloud below this sequence. S0/Sa galaxies lie between these two star populations but near the locus of the main sequence. The histograms on top of each panel in figure 8 shows the distribution of stellar masses as a function of morphological type. The stellar mass functions of S0/Sa galaxies is similar to Ellipticals whereas Spiral galaxies are on average, less massive than both Ellipticals and S0/Sa. The histograms on the right-hand side of each panel of figure 8 represents the distribution of SFRs. The S0/Sa galaxies analyzed in this work have SFRs comparable to Spirals but higher than Ellipticals but are usually more massive and so their SFRs fall in between Spirals and Ellipticals. They are usually much redder than Spirals but are more star forming that Ellipticals.

The methodology employed in the supervised approach using CyMorph and CNNs distinguishes between ETGs and LTGs and uses an overall accuracy (OA) that measures the strength of the classifications used. Reaching an OA of 98% in the TML

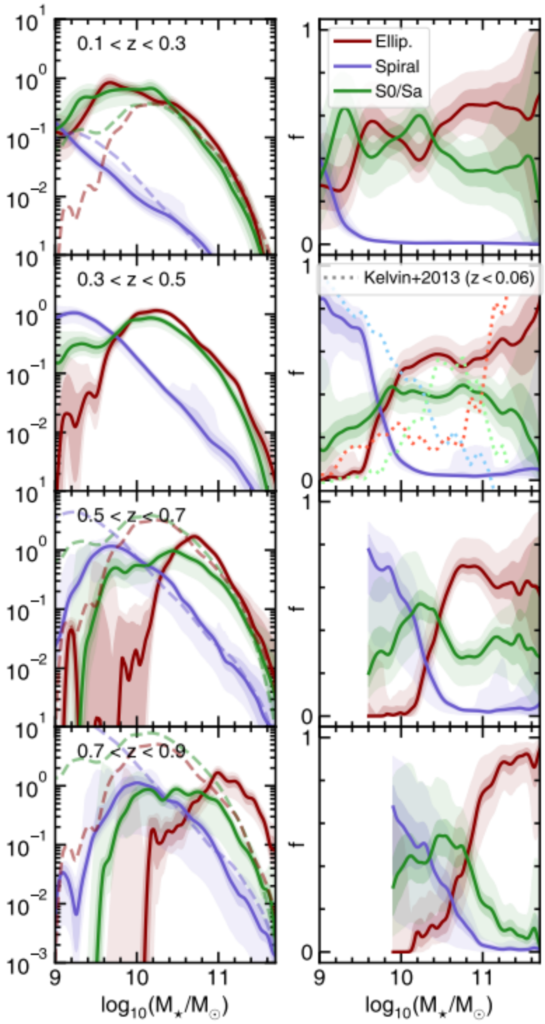


Fig. 7: **Left:** Galaxy stellar mass functions for Ellipticals (red), S0/Sa galaxies (green) and Spirals (blue) in four redshift bins. **Right:** Evolution of galaxies for 0.1 < z < 0.9. The redshift bin of 0.1 < z < 0.3 is included for completeness but can be ignored as there aren't many galaxies to analyze at this epoch. G. Martin et al. (2019)

elaborates on its efficiency. The imbalance of Spirals to Ellipticals in the training set gives a ∼ 99% OA for the Spiral dataset and a 95% in the Elliptical OA. Through GZ classes (Ellipticals and Spirals), CNN has the best approach to morphological classifications due to an OA > 94.5% in comparison to expert visual classifications Ting-Yun Cheng (2019). A class imbalance that rises from using a training set of 11 distinct morphological types reduces CNN accuracy to ∼ 65.2%. The accuracy of CNNs fall short if there is need for specific morphology's (classes) as seen in figure 9 and 10. The robustness of this methodology is tested through the reproduction of known classification trends of star formation, galaxy colour magnitudes and stellar mass as functions of morphological types at z < 1.

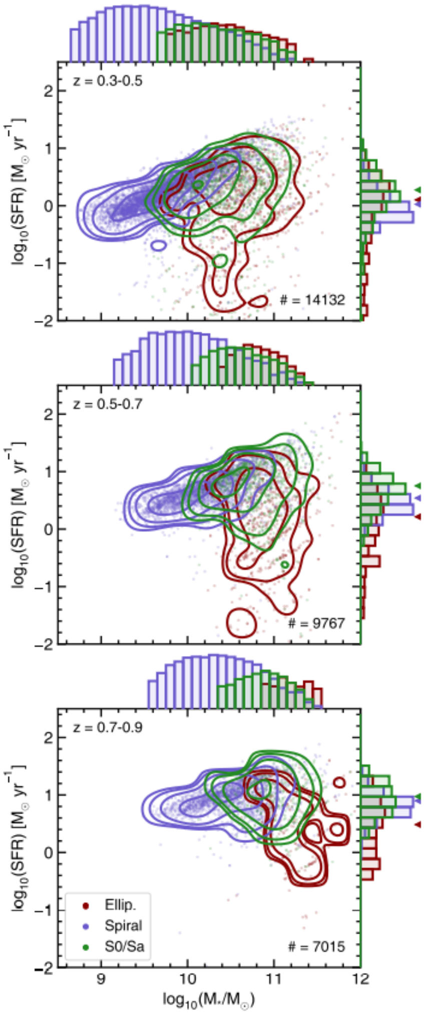
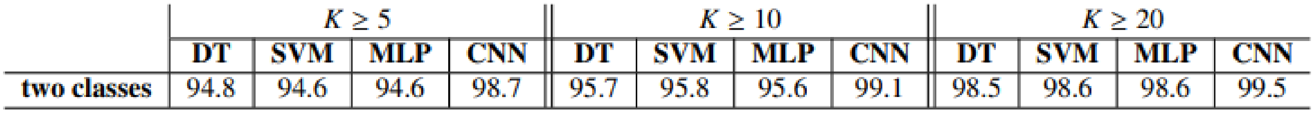


Fig. 8: Scatter and histogram plots with contours showing the distribution of galaxies as functions of SFRs and stellar mass for Elliptical (red), Spiral (blue) and S0/Sa (green) galaxies. Plots for different redshift bins show that at earlier epochs, stellar mass and SFRs were more common for Ellipical galaxies. At z < 0.5, Spirals dominate in SFR, supporting expert visual classifications. G. Martin et al. (2019)

5. Conclusions

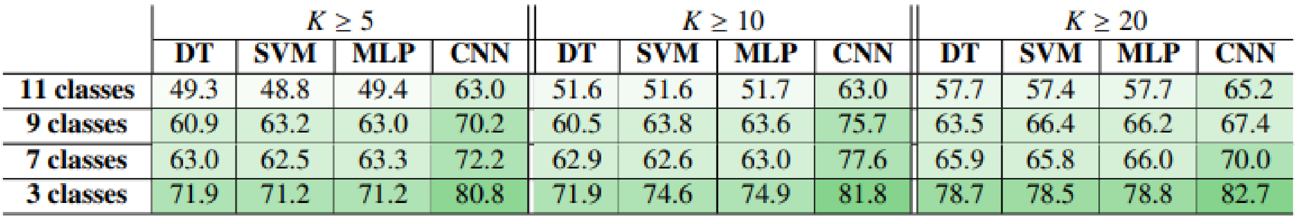
Unsupervised Machine Learning is an attractive solution to next generation big data surveys. Effective UML algorithms autonomously and rapidly compress galaxy images into subsets of morphological clusters of similar morphological characteristics. If there are less clusters (less than a hundred), the benchmarking process is flexible for visual comparisons from experts and can be used in supervised learning as pre-labelled data.

The UML explored in this paper uses k-clustering of 160 morphological clusters that extracts sub-image patches from multi-band/wavelength data that is transformed into rotationally-invariant regions of spatial information, colour and intensity. Growing neural gas and hierarchical clustering is employed to from a library of distinct patch types which assembles object feature vectors depicting the frequency for morphological char-



	$K \geq 5$				$K \geq 10$				$K \geq 20$			
	DT	**SVM**	**MLP**	**CNN**	**DT**	**SVM**	**MLP**	**CNN**	**DT**	**SVM**	**MLP**	**CNN**
two classes	94.8	94.6	94.6	98.7	95.7	95.8	95.6	99.1	98.5	98.6	98.6	99.5

Fig. 9: Overall Accuracy (OA in %) of classifications between CNN and traditional SML for two classes. K represents a parameter that divides the disk to bulge ratio of a galaxy **?** to the area of the Full Width at Half Maximum (FWHM). Larger K values have a higher OA for data from Galaxy Zoo 1. Barchi et al. (2020)



	$K \geq 5$				$K \geq 10$				$K \geq 20$			
	DT	**SVM**	**MLP**	**CNN**	**DT**	**SVM**	**MLP**	**CNN**	**DT**	**SVM**	**MLP**	**CNN**
11 classes	49.3	48.8	49.4	63.0	51.6	51.6	51.7	63.0	57.7	57.4	57.7	65.2
9 classes	60.9	63.2	63.0	70.2	60.5	63.8	63.6	75.7	63.5	66.4	66.2	67.4
7 classes	63.0	62.5	63.3	72.2	62.9	62.6	63.0	77.6	65.9	65.8	66.0	70.0
3 classes	71.9	71.2	71.2	80.8	71.9	74.6	74.9	81.8	78.7	78.5	78.8	82.7

Fig. 10: Overall Accuracy (OA) for data from Galaxy Zoo 2 classifications. Higher class imbalance (more than 3 classes) has a lower OA. The darker the green colour of the cell, the better the OA. Barchi et al. (2020)

acteristics. Testing the algorithm for galaxies in z<1 produces results from previous visual classifications on stellar mass and star formation rates for galaxies on the main sequence.

Deep Learning methods require a huge amount of data and complex computational resources to create effective classification systems. Deep learning can be hard to tune and manage and predictions require more time than UML due to the complexity of the algorithms. Traditional SML offers quite a dependable OA but slightly falls short of the CNN approach (Δ OA $\sim 4\%$). The datasets used to build SML results used Galaxy Zoo 1 and saw a similar trend to the stellar and star formation rates of the UML.

The study of galaxy morphology gives us the ability to understand the processes in which galaxies evolve over time and what mechanisms and evolutionary channels they follow. The inaccuracies of merger and star classifications seen by figure 3 shows there is a need to improve machine learning algorithms. Despite this, the similarity in results from SML and UML methods shows that combination of CNNs and clustering has the potential to create a more robust classification system. The autonomous nature of UML and the OA of CNNs are the next step in the analysis of galaxy morphology but unfortunately limited by computational complexities in administrating the methods simultaneously.

References

P.H. Barchi et al., 2020, Machine and Deep Learning Applied to Galaxy Morphology - A Comparative Study, Astronomy and Computing, 30, 23

Ting-Yun Cheng et al., 2021, Galaxy Morphological Classification Catalogue of the Dark Energy Survey Year 3 data with Convolutional Neural Networks, Monthly notices of the Royal Astronomical Society, 507, 23

J Vega-Ferrero and H Domínguez Sánchez et al., 2021,Pushing automated morphological classifications to their limits with the Dark Energy Survey, Monthly Notices of the Royal Astronomical Society, 506, Oxford University Press

G Martin and S Kaviraj and A Hocking and S C Read and J E Geach, 2019, Galaxy morphological classification in deep-wide surveys via unsupervised machine learning, Monthly Notices of the Royal Astronomical Society, 491, Oxford University Press

H Domínguez Sánchez and M Huertas-Company and M Bernardi and D Tuccillo and J L Fischer, 2018, Monthly Notices of the Royal Astronomical Society, 476, Oxford University Press

de Vaucouleurs, G., 1963, Revised Classification of 1500 Bright Galaxies, Astrophysical Journal Supplement, 8, 31

E.P. Hubble, Realm of the Nebulae, New Haven: Yale University Press, 1936.

Lintott et al., Jan, 2011, Monthly Notices of the Royal Astronomical Society, Volume 410, Issue 1, pp. 166-178

E. L. Nielsen et al. The Gemini Planet Imager Exoplanet Survey: Giant Planet and Brown Dwarf Demographics From 10-100 AU, 2019, Earth and Planetary Astrophysics

M. Reza, 2021, Galaxy morphology classification using automated machine learning, 2021, Astronomy and Computing, 37

Galaxies, human eyes, and artificial neural networks, 1995, Science. 1995

An automatic taxonomy of galaxy morphology using unsupervised, 2017, MNRAS 473

A Growing Neural Gas Network Learns Topologies, 1995, Advances in Neural Information Processing Systems