Automating Galaxy Classification with Unsupervised Machine Learning

Kirsten Elliott (MSc) [1]

Supervised by Assoc. Prof. Michelle Lochner [1]

[1] Department of Physics and Astronomy, University of the Western Cape, Bellville, Cape Town, 7535, South Africa

UNIVERSITY *of the* WESTERN CAPE

NASSP NATIONAL ASTROPHYSICS AND SPACE SCIENCE PROGRAMME

NRF National Research Foundation

science, technology & innovation
Department: Science, Technology and Innovation
REPUBLIC OF SOUTH AFRICA

1 Why automate classification?

A galaxy's morphology encodes valuable information about its formation, merger and star-formation history. Citizen-science projects such as Galaxy Zoo have produced excellent labelled catalogues, but upcoming surveys (e.g., the Vera C. Rubin Observatory's LSST) will produce enormous imaging datasets, ultimately observing billions of galaxies over its ten-year survey, far beyond what volunteers can label by eye.

This thesis investigates whether **unsupervised clustering** of **self-supervised features** can recover meaningful **morphological classes** without the use of labels.

2 Data: Galaxy Zoo DECaLS (GZD)

GZD-5 pairs DESI Legacy Imaging Survey cutouts with detailed volunteer morphology votes for over 300 000 galaxies (Walmsley, et al. 2022). The votes are collected across a set of questions about the presence of a particular morphological feature, which forms a decision tree of increasing complexity.

The **volunteer vote catalogue** (253 286 galaxies) as well as the **image catalogue** (269 760 galaxies) was released in Walmsley et al. (2022), and made available on Zenodo.

3 Building on self-supervised features

Mohale and Lochner (2024) implemented a feature extraction algorithm built using a **ResNet-18 backbone** (He et al. 2016) to the GZD-5 images. Through visual inspection of the feature space, a set of images (39 185 galaxies) showing prominent artifacts were removed. The remaining feature set (230 575 galaxies) was then fine-tuned with the self-supervised method **Bootstrap Your Own Latent (BYOL)** (Grill et al. 2020). This high dimensional feature space (512 dimensions) was then reduced to 29 dimensions using Principal Component Analysis (PCA).

This thesis makes direct use of this 29-D feature space of 230 575 galaxies, applying two unsupervised clustering algorithms to probe its structure.

4 Two clustering algorithms

Both the **Bayesian Gaussian Mixture Model (BGMM) and Mini-Batch K-Means (MBKM)** (Pedregosa et al. 2011) algorithms are used to cluster the same 29-D feature space of 230 575 galaxies. BGMM was used by Mohale and Lochner (2024) to separate this feature space into 20 clusters; MBKM is introduced here as a faster alternative.

BGMM
BayesianGaussianMixture
Mixture of full-covariance Gaussians under a Dirichlet-process prior, which automatically suppresses unnecessary components.

n_components	20
weight_concentration_prior	0.5
n_init	10
max_iter	1000

MBKM
MiniBatchKMeans
Partitions the feature space into exactly k-clusters using mini-batches

n_clusters	20
n_init	10
max_iter	1000

5 Confidently-Labelled Test Set

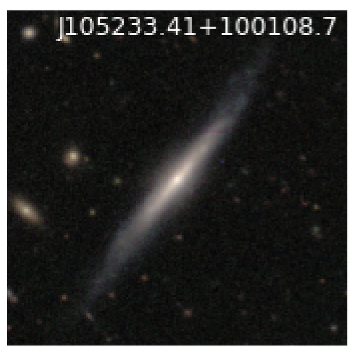
To evaluate the unsupervised clusters, a smaller subset of the 230 575 galaxies is created which all have labels based on high volunteer agreement. The volunteer vote catalogue was refined through a number of vote-count cuts (applied to specific questions in the GZD-5 decision tree) to create the **test set** of 15 721 galaxies with confidently voted morphological class labels.

Volunteer vote catalogue	253 286
Excl wrong_size warnings	250 930
Excl. < 34 total votes for Q1	57 266
Excl. < 50% of Q1 total votes for Q2/Q3/Q5	47 634
Excl. rare tage combinations (< 100 members)	47 153
Isolate 3 morphological classes	24 858
Cross-match with PCA features	15 721

Round Elliptical
5 780

Spiral
8 358

Edge-On
1 583



6 Evaluating cluster quality

Once each algorithm partitions the feature space, every cluster is assigned the **majority morphological class label** among its labelled members.

A 2-D **Uniform Manifold Approximation and Projection (UMAP)** plot (McInnes et al. 2018) of the 29-D feature space is created in order to visually inspect the feature space. In order to visualize how successfully the clustering algorithm separated the three morphological classes, each galaxy is coloured by its predicted morphological class.

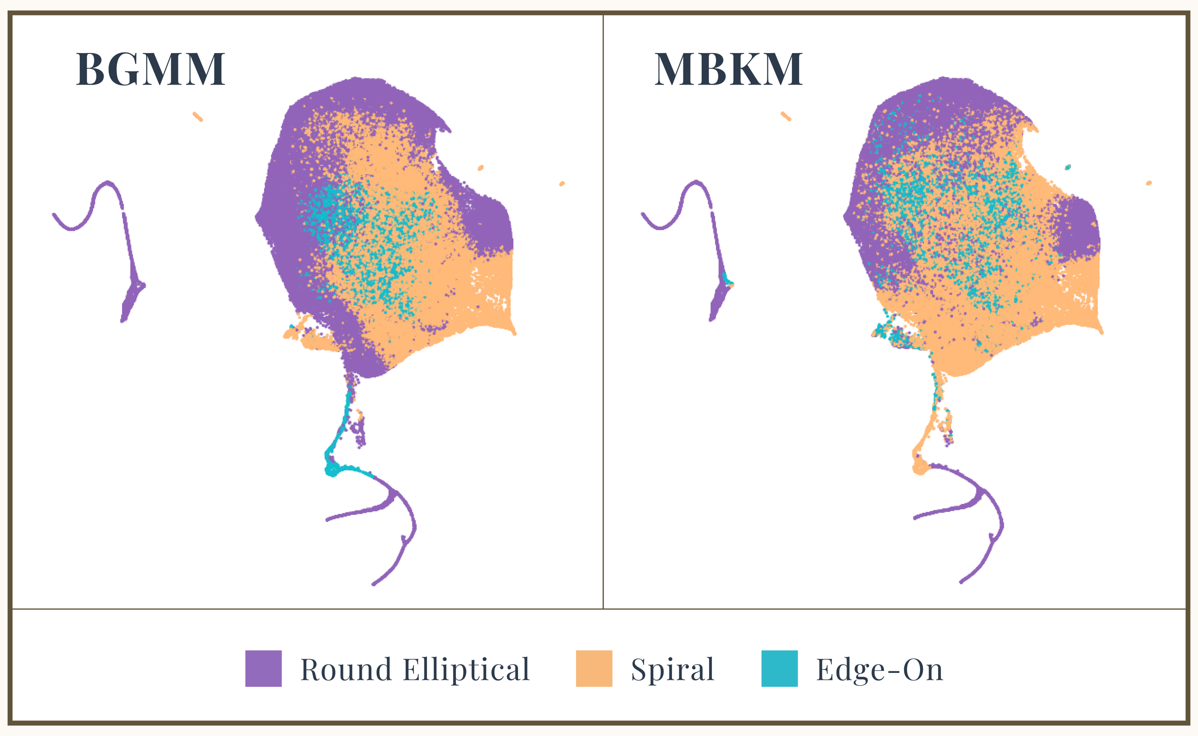


Figure 1: UMAP projections (using identical random seeds for comparability) of the 29-D feature space of all 230 575 galaxies, clustered using the BGMM (left) and the MBKM (right) algorithms. Each point represents one galaxy projected from the original 29-dimensional feature space into two dimensions and is coloured according to the dominant morphological class of its assigned cluster.

The **accuracy** achieved by the clustering algorithm is calculated as the fraction of test set galaxies whose cluster label matches their confident morphological class label.

Table 1: Classification accuracy of morphological classes predicted from cluster assignments produced by the BGMM and MBKM algorithms, evaluated on the test set of 15,721 galaxies.

Morphological Class	No. of Galaxies	Classification Accuracy (%)	
		BGMM	MBKM
Round Elliptical	5 780	73.0	70.1
Spiral	8 358	82.5	78.8
Edge-On	1 583	64.0	70.4
Overall	15 721	77.2	74.7

7 Conclusions and Outlook

- Both BGMM (**77.2% overall**) and MBKM (**74.7% overall**) recover broad galaxy morphology classification from the same self-supervised features at comparable overall accuracy.
- Spirals achieve the highest per-class accuracy (BGMM: 82.5% , MBKM: 78.8%), possibly due to their visually distinctive features.
- UMAP projections coloured by predicted morphological class show very similar class separations. Edge-On galaxies are physically a viewing-angle subset of spirals rather than a distinct morphological family, so their features overlap substantially with the Spiral class in the 29-D feature space

MBKM (which runs in under 5 seconds) achieves comparable overall accuracy to BGMM (~ 6hrs) at substantially lower computational time, making it the more practical choice for scaling to larger, multi-survey datasets.

Future work: Quantifying per-cluster uncertainty and testing whether it correlates with volunteer disagreement in the Galaxy Zoo vote fractions.

Prior-Work

Mohale, K. and M. Lochner (2024). Enabling Unsupervised Discovery in Astronomical Images through Self-Supervised Representations. *MNRAS*. 530(1), 1274–1295.

Data

Walmsley, M. et al. (2022). Galaxy Zoo DECaLS: Detailed Visual Morphology Measurements from Volunteers and Deep Learning for 314 000 Galaxies. *MNRAS* 509(3), 3966–3988.

Self-Supervised Learning

Grill, J.-B. et al. (2020). Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. *NeurIPS*. 33, 21271–21284.

ResNet Backbone

He, Kaiming, et al. (2016). Deep residual learning for image recognition. *CVPR* 2016. 770–778

Clustering Algorithms

Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python. *JMLR*. 12, 2825–2830

Visualization

McInnes, L., Healy, J and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection. *arXiv:1802.03426v3*