Galaxy Classification using Machine Learning Methods

William Kwan
University of Toronto

Abstract
Galaxy classification is a cornerstone of extragalactic astronomy, providing crucial insights into the formation and evolution of the universe. In this study, we review current research in galaxy classification, focusing on the terminology, morphological types, and the role of redshift in shaping observable characteristics. We highlight the importance of citizen science initiatives such as the Galaxy Zoo 2 project, which leverages public participation to classify galaxies based on their morphological features. Using the GZ2 votes, we employ machine learning techniques to automate galaxy classification. Starting with a Linear Regression model using extracted features, we transition to a neural network, then a convolutional neural network on the galaxy images themselves, and finally finetuning Zoobot. Our results demonstrate that the CNN is able to best classify galaxies, however, other models are very close in performance, with over 70% test accuracy using 61k galaxy images. This work is a very brief and hands on introduction to the field of statistical learning with astronomical images.

Contents

1 Introduction

First, we present a background overview of the differences in galaxy structure over cosmic time. This gives us insight into the purpose of creating an automated algorithm to classify galaxies. The progress made in identifying differences in galaxy structure allows us to implement machine learning methods in the GZ2 project.

The Introduction gives an overview of the terminology used, including the types of galaxies, their dominance at different redshifts, and how these types of galaxies are identified using features.

1.1 Evolution of Galaxy Structure

The most significant finding in galaxy evolution is that distant galaxies are more compact than galaxies of the same mass in the local universe. For example, the effective radii increase by up to a factor of 5 for today's galaxies versus galaxies at $z = 3$ for the same stellar mass.

Visual classification of galaxies can predominantly be put into a few classes, mainly: spirals, ellipticals, and irregulars/peculiars. Ellipticals are smooth, featureless galaxies that have little gas and dust, and have minimal star formation. They often have larger stellar masses, as most of the massive galaxies in the nearby universe are elliptical galaxies. It appears that on average, it is often the case that galaxies with a higher central concentration have larger stellar mass, and thus have significant and concentrated bulges. Sersic demonstrated most massive ellipticals would follow a smooth de Vaucouleurs profile.

Spiral galaxies are less massive and bluer and have evidence for ongoing star formation. Spiral galaxies are often classified as barred or unbarred galaxies. Early studies found that the bar fraction evolves significantly, from $z = 0.84$ to $z = 0.2$, we can see that bar fraction increases from 20% to 80% of all disk galaxies, while the bar fraction for redder galaxies are mostly constant.[1]

2 Methodology

Overview of Feature Extraction and ML Classification Algorithms

2.1 Feature Extraction

The most traditional approach to understanding galaxy structure involves classifying galaxies by their visual features. Using quantitative methods, we are able to create precise measurements from a galaxy's visual appearance and translate it into features in which we can further analyze. There are parametric and non-parametric approaches to feature extraction.

The main parametric method is the Sersic profile, which has the form:

$$I(R) = I_0 \exp(-kR^{1/n})$$

where I_0 is the light intensity at $R = 0$. The Sersic index, n, quantifies controls how "curved" the profile is. Higher values indicate a higher central concentration, and smaller values indicate less central concentration. For example, elliptical galaxies has $n = 4$, while spiral galaxies typically have around $n = 1$.

For non parametric methods, we mainly use **CAS**, where **C** measures the ratio of light in a galaxy's central region to its outer regions. It has a strong correlation to the Sersic Index, which also measure light concentration in a galaxy. The inner and outer radii defined using the total amount of light given some Petrosian radius. **A** measures the asymmetry index, which is how asymmetric a galaxy is after rotating it along the center axis by 180 degrees. This is quantified using the formula:

$$A = \min\left(\frac{\Sigma|I_0 - I_{180}|}{\Sigma|I_0|}\right) - \min\left(\frac{\Sigma|B_0 - B_{180}|}{\Sigma|I_0|}\right)$$

where I_0 is the original galaxy image, and I_{180} is the image after rotating it from its center by 180 degrees. Here, there are other computational steps involving using a blank background B_0 to find its center of rotation. Typical asymmetry values for local galaxies are: ellipticals around $A \sim 0 - 0.4$, spirals around $A \sim 0.07 - 0.2$, mergers around $A \sim 0.12 - 0.51$, and starbursts around $A \sim 0.31 - 0.74$. **S** represents clumpiness or smoothness describes the fraction of light in a galaxy that is contained in clumps. For example, clumpy galaxies contain a large amount of its light in high spatial frequencies (clumps), whereas smooth galaxies contain its light in low spatial frequencies. Galaxies undergoing heavy star formation have higher clumpiness values. **S** is

measured by:

$$S = 10\times\left[\left(\frac{\Sigma(I_{x,y} - I^{\sigma}_{x,y})}{\Sigma I_{x,y}}\right) - \left(\frac{\Sigma(B_{x,y} - B^{\sigma}_{x,y})}{\Sigma I_{x,y}}\right)\right]$$

where the original image $I_{x,y}$ is blurred to create a secondary image $I^{\sigma}_{x,y}$, which is then subtracted from the original image to locate high frequency areas. The smoothing value σ is determined by a function of the radius of the galaxy.

Lastly, we also employ the Gini/M20 parameters. These parameters do not involving subtracting a part of the galaxy to determine asymmetry and clumpiness. In principle, this may be less sensitive if we have a noisy background. For the Gini index, higher values indicate a very unequal distribution, such as all the light being concentrated in one pixel, whereas a lower value gives a more even distribution. In this case, a galaxy is an image with n pixels, each with a flux f_i. The Gini index is measured by:

$$G = \frac{1}{|f|n(n-1)}\sum_{i}^{n}(2i - n - 1)|f_i|$$

where $\bar{f}$ is the average flux. The M20 parameter value is similar to **C**, the concentration in **CAS**, except a high concentration M20 value implies that the concentration can be anywhere, not just the center. The M20 value is the moment of the fluxes of the brightest 20% of the galaxy, normalized by the total light moment for all pixels. The M20 is given by:

$$M_{20} = \log_{10}(\frac{\sum_i M_i}{M_{\text{tot}}})$$

where the value of M_{tot} is:

$$M_{\text{tot}} = \sum_{i}^{n} M_i = \sum_{i}^{n} f_i\left[(x_i - x_c)^2 + (y_i - y_c)^2\right]$$

Although in literature, other parameters such as the multiplicity index, multimode, or deviation is used, we did not consider them.

2.2 Overview of ML Algorithms

In this report, we considered many machine learning approaches, including traditional machine learning algorithms, and using convoluted neural networks. To begin, we first considered using simple linear regression with one feature as benchmarks. These served as the minimal "target" in which our more complex models should perform better than. To extract features such as the ones described in the previous section, the `statmorph` library is used, namely by extracting one channel of the image and convoluting with a Gaussian PSF. Those features can then be used for logistic regression algorithms, or neural networks. `scikit-learn` and `pytorch` libraries are used for these algorithms.

The validate our performance, we first use CV (cross-validation) to split the dataset into training-validation-testing. We first split the dataset into 80-20 for train and test, respectively, and another 80-20 on the training dataset for validation. As for the metric, the RMSE (root mean squared error) function is used:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_i - a_i)^2}$$

where p_i is the predicted probability (or vote count) from our model, and a_i is the actual. Further insight into model architectures will be put in Section 3.

3 Results on Classification and Discussion

Results and Statistical Inference

3.1 Data Overview

The data is collected from the Kaggle Galaxy Zoo Challenge, which includes images for 61578 galaxies. Probability distributions for the classifications of the galaxy images are also provided. The probabilities are vote proportions taken from the **GalaxyZoo** project, where participants are given a galaxy image and have to classify given a decision tree. While the specifics of each question won't be delved on here, the classification begins with general questions, such as smooth, spiral or irregulars, and narrows to specifics such as the number of spiral arms or the type of irregularity. The probabilities initially sum to 1.0 at each question, but are weighted cumulatively based on prior probabilities. As such, during the inference stage, we will focus on top level questions. An overview of the types of galaxies are displayed in the Appendix (Figure 5).

3.2 Exploratory Data Analysis

With data from the GZ2 paper, we put figures of different proportions of galaxies in each question (i.e. edge on vs not) as a function of redshift in the Appendix (Figure 4). Note that each proportion is a proportion of the previous question. For example, the Edge-On proportions are from the "Featured or Disk-Yes" voters from the previous question, not a proportion of the entire dataset. However, the "Smooth or features" question is a proportion of the entire dataset as it is the first question.

3.3 Benchmarking

As this is a classification problem, we need to have a set baseline benchmark to compare our models against. In the Galaxy Zoo original paper, the machine learning model used had an over 90% classification accuracy on predicting smooth versus disk versus star images[2]. Although the dataset used in the paper is larger than the one on Kaggle, it still serves as a good baseline for good performance.

We also tested extremely simple models: the average pixel model, and the center pixel model. These single feature models used linear regression and was able to output RMSE values of **0.1599** and **0.1564**, respectively. The main goal of subsequent models is to accurately classify smooth versus disk galaxies, as well as providing a good RMSE score.

3.4 Statmorph Models

We use Statmorph to extract the features discussed in the introduction, mainly **CAS**, Gini, M20, and Sersic Model index. This is fit using a Linear Regression model on the 20% of the splitted test data. A neural network, using 3 hidden layers (44, 88, 22) is also used.

With the 20% test set, we are able to achieve both higher than 70% accuracy in classifying smooth versus disk galaxies. The RMSE score is **0.1302** and **0.1482** for the Linear Regression model and the neural network model, respectively. While there is some class imbalance, 42% for smooth vs 55% for disk galaxies, we're able to achieve **78%** accuracy and **70%** accuracy for classifying smooth vs disk galaxies using LR and NNs, respectively. As there are more disk galaxies (edge on or spirals) than smooth galaxies, it is not surprising that we have a better accuracy to detecting disk galaxies.

3.5 CNN Model

For the CNN, we first downscale the image, from 424x424 to 128x128 in order to adjust for training times. The image is a cropped version of the full image. We then use 3 convolutional layers (64, 96, 128) with batch normalization and flattening. Then the output from the convolution layers is put into 2 fully connected layers with 15% dropout for classification using sigmoid as the activation function. The result gives **84%** accuracy to detecting smooth vs disk galaxies, and a RMSE of **0.1134**.

3.6 Differences between TML and DL

We see that the two approaches (using statmorph vs using a CNN) gives very similar results with accuracy. However, the losses differ considerably. We note that this can be from a misspecification of the Gaussian PSF for statmorph, which gives rather inaccurate feature extraction. We can also attempt to explain what causes misclassification at the first question. Figure 1 shows some examples of mis-classification per model. Some misclassified examples include images with another bright light source obstructing the galaxy, images with more than one galaxy contained in them, as well as images with galaxies at are fainter than usual. This reinforces the importance of treating images that are different from the average image separate as a specific set. We also note that non-parametric models such as the CNN must always rely on a non obstructed image to get ideal performance.

We also use the AUC to measure performance. The ROC is the probability curve and the higher the AUC, the better the model is at predicting smooth galaxies as smooth galaxies and disk galaxies as disk galaxies.

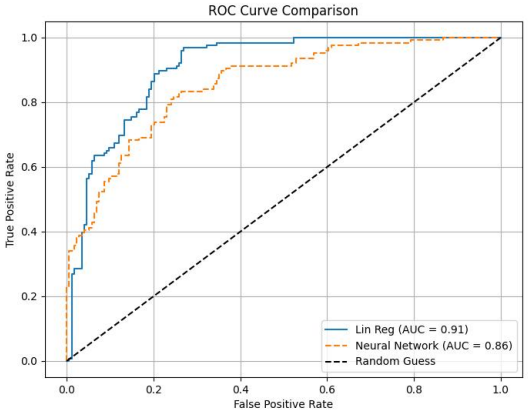

Figure 1: Misclassified examples from the CNN model

4 Analzying Model Weights and Zoobot

Extracting feature weights, comparison with Zoobot performance

4.1 Model Weights

Representations provide us the opportunity to look at similar galaxies in the feature space. We can extract model weights for galaxies, and list its closest neighbors to check if our model is giving weights as it should have. We note that identifying nearest neighbors in the representation space of $D = 8192$ is extremely computationally intensive. We reduce the dimensionality using `sklearn`'s `IncrementalPCA` down to $D = 15$ to preserve variance while making the computation efficient. Using the distance metric $\Sigma_i |p_i - q_i|$, we are able to compute the nearest neighbors of each galaxy in the representation space of $D = 15$. The representations can also be visualized using `umap` after using Incremental PCA. This gives us a 2D representation, in which we are able to group galaxies according to the top level questions. Figure 2 shows representation maps for Smooth vs Disk galaxies as in the first question, and the image on the right further splits these disk galaxies into neither (oddities), spirals, and edge on galaxies.

Figure 6 and 7 in the appendix also shows 5 spiral galaxies and 5 smooth galaxies and their closest neighbours. We see that in the majority of cases, the closest neighbours are classified as the same type. We also note a limitation of our model, where if we classify galaxies as an oddity, for example, a merger, its nearest neighbor will not be a merger in the representation space. This is most likely a result of our model

4.2 Zoobot Implementation

We also implement Zoobot, a pre-trained model in which we can fine-tune our own data to fit our classification task. Zoobot is created by Mike Walmsley[4], and it uses pretrained galaxy images model to solve any galaxy morphology problem, as long as we have a defined training set. Here, we use the ConvNeXT-Nano model, and provide it with a 80/20 train-test split, with only the first GZ2 question (smooth, disk, or artifact). The Zoobot model is able to achieve a 83% test accuracy, very similar to the 84% test accuracy achieved by our own CNN. Zoobot has a way to extract learned representations, in which we use the same IPCA-UMAP procedure to generate our representation map (Figure 3). We also include galaxies and its nearest neighbours (as we did with the CNN) in Figure 8.

4.3 Zoobot Performance Comparison

Firstly we note that the representation map does not appear to separate smooth vs disk galaxies, even though we have a very similar classification rate to the CNN. We think that this maybe because UMAP reducing representations down to 2 dimensions does not accurately move the decision boundary to 2 dimensions. However, this remains uncertain as with the Zoobot paper it is able to do so accurately[4]. In Figure 8, we see this problem again. Nearest neighbours does not necessarily involve a galaxy similar visually as it did with the CNN. For example, smooth galaxies having spirals or edge on galaxies as its closest neighbor. We don't believe this is a problem with classification, as we are able to get 83% accuracy, but rather with the dimensionality reduction tool itself (IPCA or UMAP).

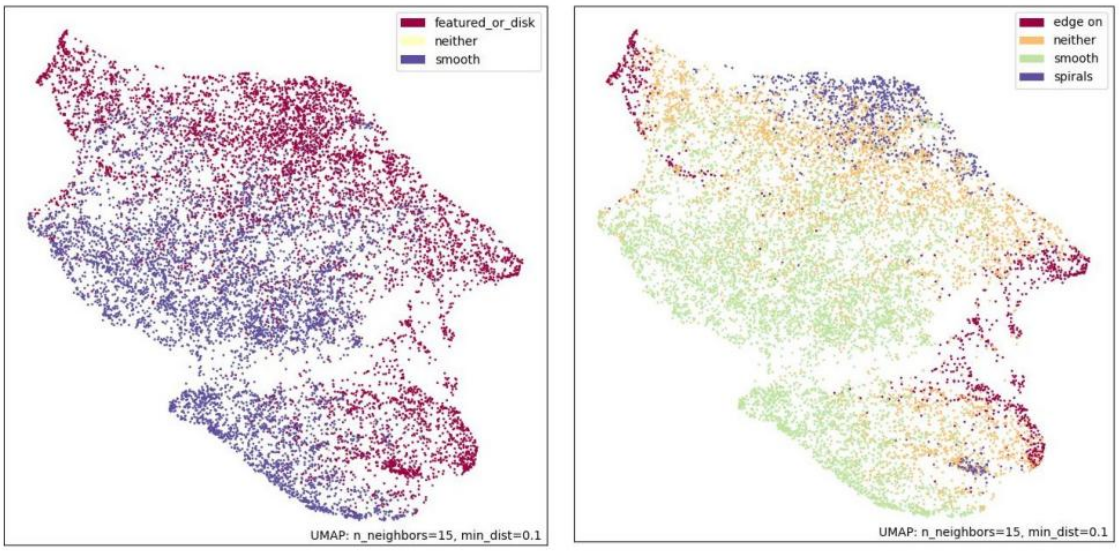

Figure 2: UMAP Representations for Smooth vs Disk galaxies. Map on the right further divides the disk galaxies into edge on, neither and spirals

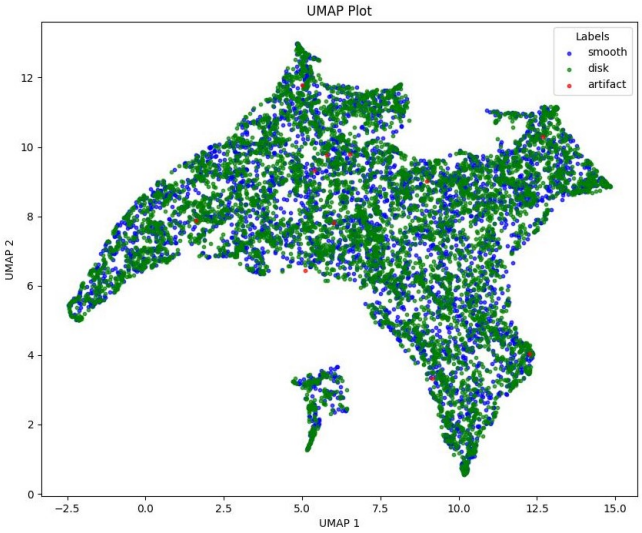

Figure 3: UMAP Representations for Smooth vs Disk galaxies using Zoobot

5 Conclusions

With more surveys and astronomical data coming up, it is important that galaxy morphology classification can be done with accuracy and efficiency. We presented different types of models to accomplish these goals. This is one of the first steps in which we can use to further study our universe. We investigated TML, DL, and CNN models and compared their performances. We saw the limitations of each model, such as the feature extraction models having worse accuracy than the CNN models, the CNN model misclassifying when the image is obstructed, or the Zoobot model not being able to produce a separable representation map. We also looked at different features, such as C, A, S, Gini, and M20, and how they are implemented in Statmorph. To summarize using first question test accuracy, we are able to achieve 78% using a LR model with feature extraction, 70% using a NN model with feature extraction, 84% using a CNN, and 83% using Zoobot. We believe that with more data (more than 10x the Kaggle dataset), and with more powerful processing units to train larger models, we will get better accuracy scores. Overall, this project serves as a good introduction to the field of using Machine Learning to classify astronomical data, and the Galaxy Zoo project.

Thank you Prof Antonio for the supervision with this project!

References

[1] Christopher J. Conselice. The evolution of galaxy structure over cosmic time. *Annual Review of Astronomy and Astrophysics*, 52:291–337, 2014.

[2] Chris J. Lintott, Kevin Schawinski, Anže Slosar, Kate Land, Steven Bamford, Daniel Thomas, M. Jordan Raddick, Robert C. Nichol, Alexander Szalay, Dan Andreescu, Phil Murray, and Jan Vandenberg. Galaxy zoo: Morphologies derived from visual inspection of galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 389:1179–1189, 2008.

[3] C. E. Moody, A. J. Romanowsky, T. J. Cox, G. S. Novak, and J. R. Primack. Radial trends

in the intrinsic shapes and kinematics of simulated early-type galaxies. *Monthly Notices of the Royal Astronomical Society*, 435(4):2921–2933, Nov 2013.

[4] Mike Walmsley, Micah Bowles, Anna M. M. Scaife, Jason Shingirai Makechemu, Alexander J. Gordon, Annette M. N. Ferguson, Robert G. Mann, James Pearson, Jürgen J. Popp, Jo Bovy, Josh Speagle, Hugh Dickinson, Lucy Fortson, Tobias Géron, Sandor Kruk, Chris J. Lintott, Kameswara Mantha, Devina Mohan, David O'Ryan, and Inigo V. Slijepevic. Scaling laws for galaxy images, 2024.

5.1 Appendix

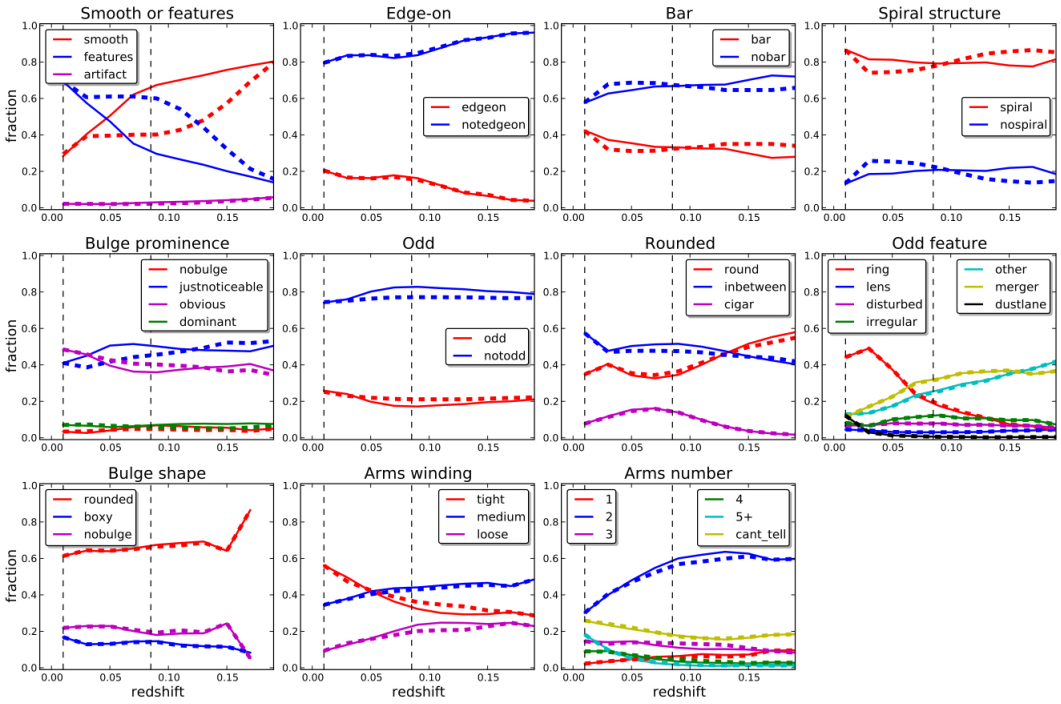

Figure 4: Type fractions as a function of redshift for GZ2. Solid lines show vote fractions, thick lines show debiased vote fractions. From GZ2 paper.[3]

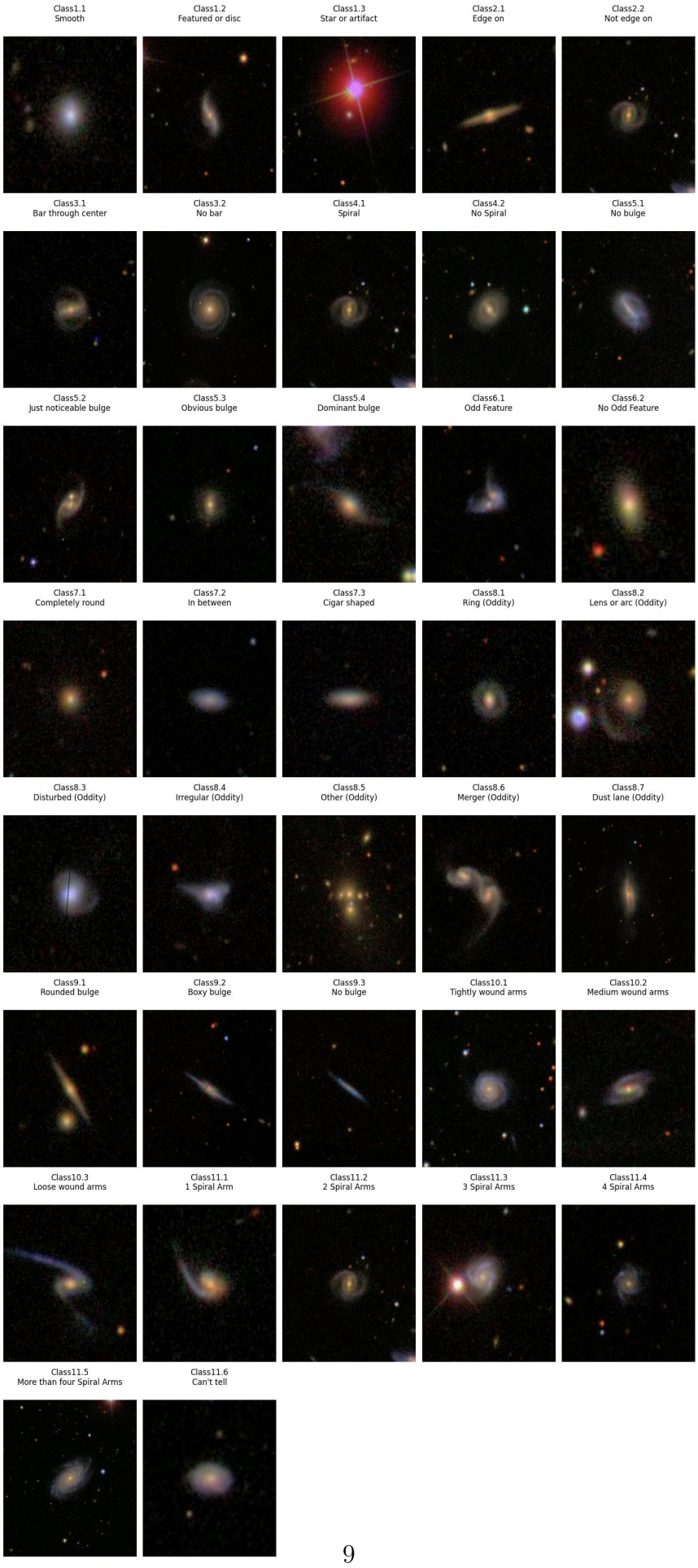

Figure 5: Highest Probability Image of each class.

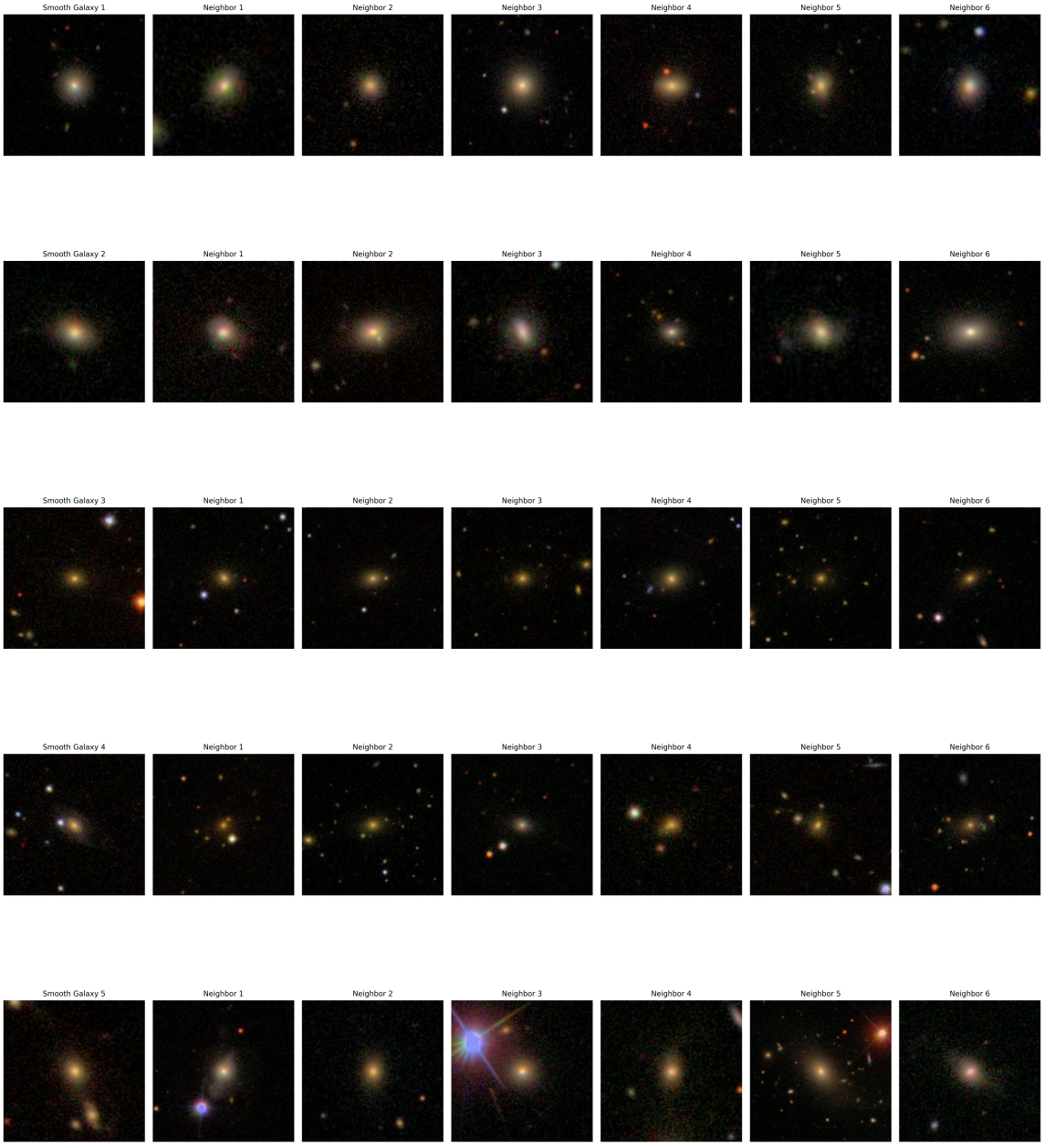

Figure 6: Smooth galaxies and closest neighbours in the UMAP representation

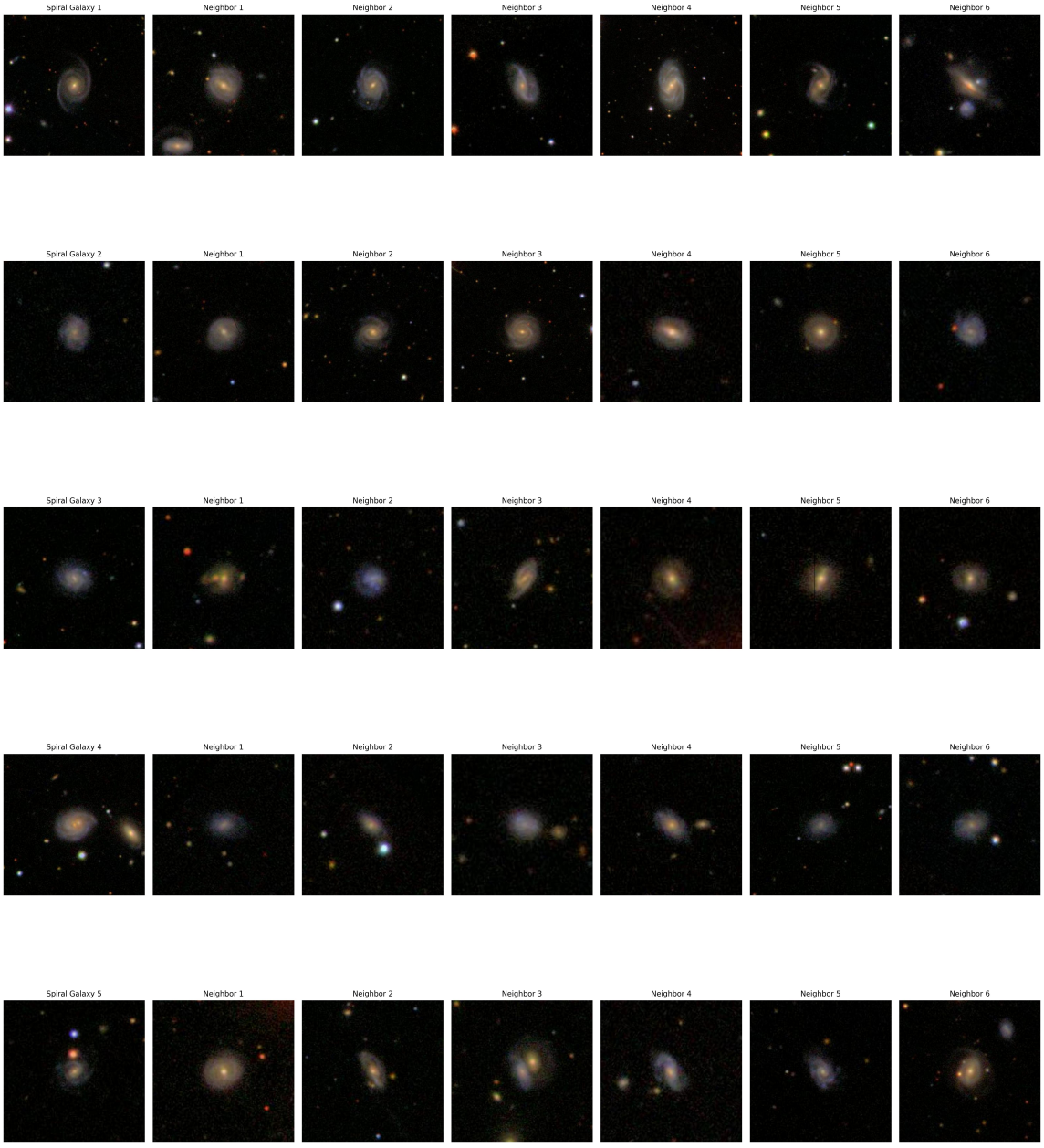

Figure 7: Spiral galaxies and closest neighbours in the UMAP representation

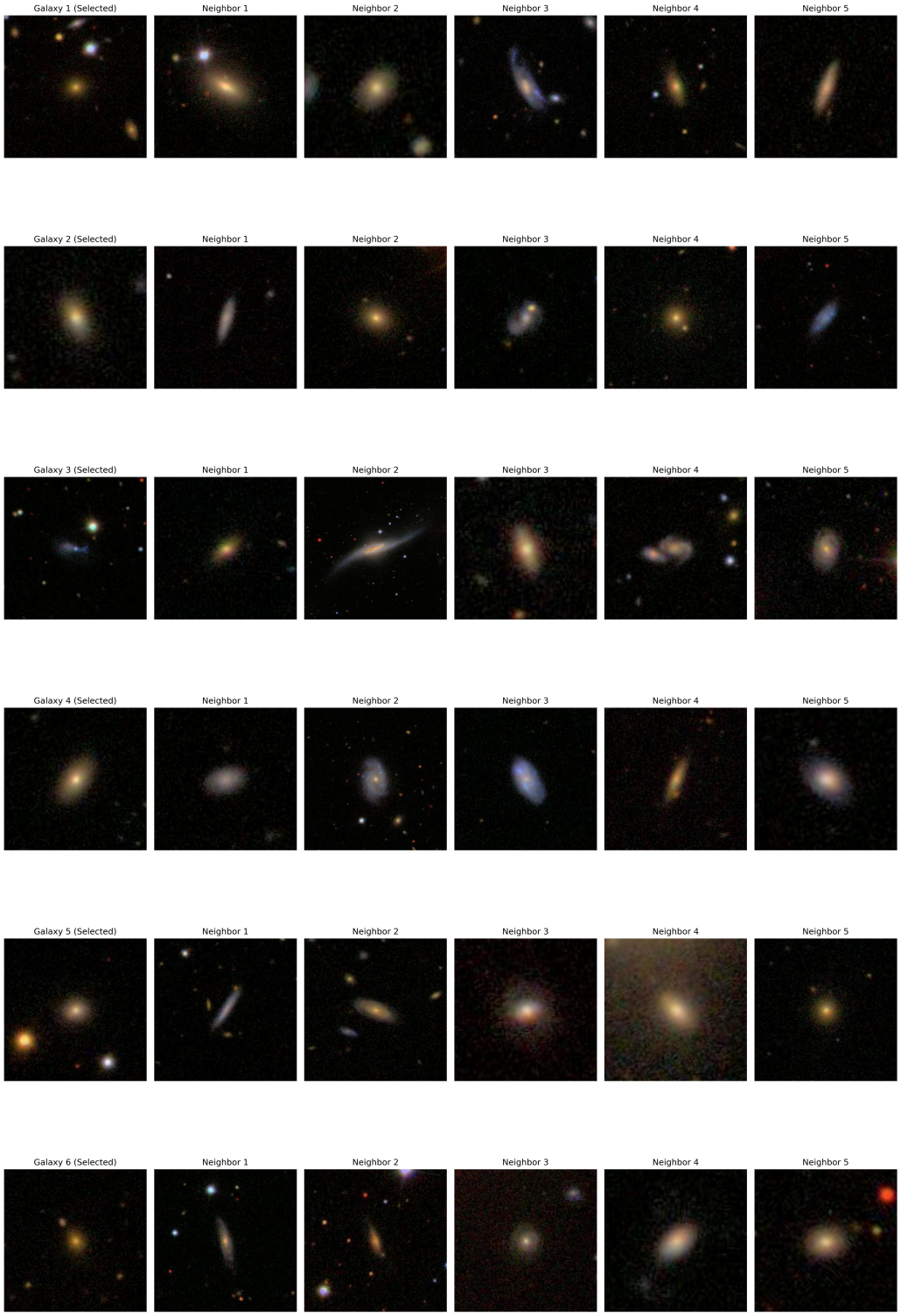

Figure 8: Zoobot randomly selected galaxies and its closest neighbours