Final Project Report
Galaxy Classification with Random Forest Classifier

I. Description

The Galaxy Zoo project brought citizen scientists to help with classifying millions of galaxy images from the Sloan Digital Sky Survey. I used a small subset (about 23,000) of galaxy images that received majority voting in the Galaxy Zoo project to train a random forest classifier.

II. Method

I used the OpenCV package to process raw optical galaxy images and identify the basic shape of the galaxy using contours. Then, I extracted different geometric and photometric features such as eccentricity, form factor (area per perimeter squared), and light intensity profiles (galaxy_features_extraction.py). I then built a random forest classifier and ran hyperparameter tuning (parameter_tuning.py). I tested the random forest classifier on categorizing galaxy images into ten groups as indicated in figure 1 (galaxy_analysis.py). Finally, I analyzed the predictive performance and the important features on each classification layer (other Python files).

III. Results & Analysis

For the general classification into all 10 galaxy types, a single decision tree has an accuracy of about 56%. A random forest with 50 or more decision trees and other fine-tuned parameters (figure 3-5) improves the accuracy to about 69% (figure 2). The overall most important features are eccentricity, form factor, Gini coefficient, intensity differences between each color band, and central (10x10 pixel) intensity (figure 6).

The random forest can distinguish between smooth and disk galaxies very well with an 85% accuracy (figure 8). The histograms of the Gini coefficient and the intensity difference between r and i bands separate themselves clearly among smooth and disk galaxies, so the photometric features play the strongest role in this classification (figure 9-11).

Smooth galaxies are further classified into 3 groups (completely round, in-between, and cigar-shaped) with an 89% accuracy (figure 12). The roundedness of a galaxy is highly correlated with the eccentricity and the form factor (figure 13-15). Similarly, disk galaxies are further classified into 2 groups (edge-on and face-on) with a 96% accuracy (figure 16). Galaxies viewed edge-on are more eccentric with lower form factors (figure 18-19).

Disk and edge-on galaxies are further classified into 3 groups based on their central bulge (round, boxy, and no bulge) with an 89% accuracy (figure 20). Here, the central intensity and the central intensity differences between color bands play the strongest roles because they are indications of the bulge (figure 21-23). The classifier struggles to predict boxy-bulge galaxies correctly because this group only accounts for less than 1% among all disk and edge-on galaxies.

Disk and face-on galaxies are further classified into 4 groups based on their spiral arms (no, tight, medium, and loose spiral arms) with a 62% accuracy (figure 24). Here, the intensity differences between the r and i bands can distinguish between no spiral arm and other subgroups (figure 26). The asymmetry factor can further distinguish loose spiral galaxies apart from the others (figure 27). In this classification layer, there is no feature that is as strongly correlated with the subgroups compared to the other layers, so the classifier performs relatively worse. Also, because no-spiral galaxies account for nearly 60% of all disk and face-on galaxies, the classifier is biased towards predicting that an unknown galaxy has no spiral arms.

IV. What I Learned from this Project

This project expanded my machine learning experience through processing raw images, extracting features, investigating the importances of different features, assembling decision trees into random forests, and fine-tuning parameters. I also learned how to investigate different classification layers and their corresponding important features through confusion matrices and histograms. In general, I have understood the inner workings of decision trees and random forests better. Computational skills aside, I also gained extra understanding into the photometric and geometric features of galaxies. The intuition I had from classifying galaxy images has helped me understand Hubble's tuning-fork classification scheme more thoroughly. Finally, my overall academic reading comprehension and research skills have improved from reading multiple papers about galaxy classification, looking for relevant information that can benefit my project, and browsing through coding documentations and tutorials.

Appendix

I. Galaxy Classification Scheme

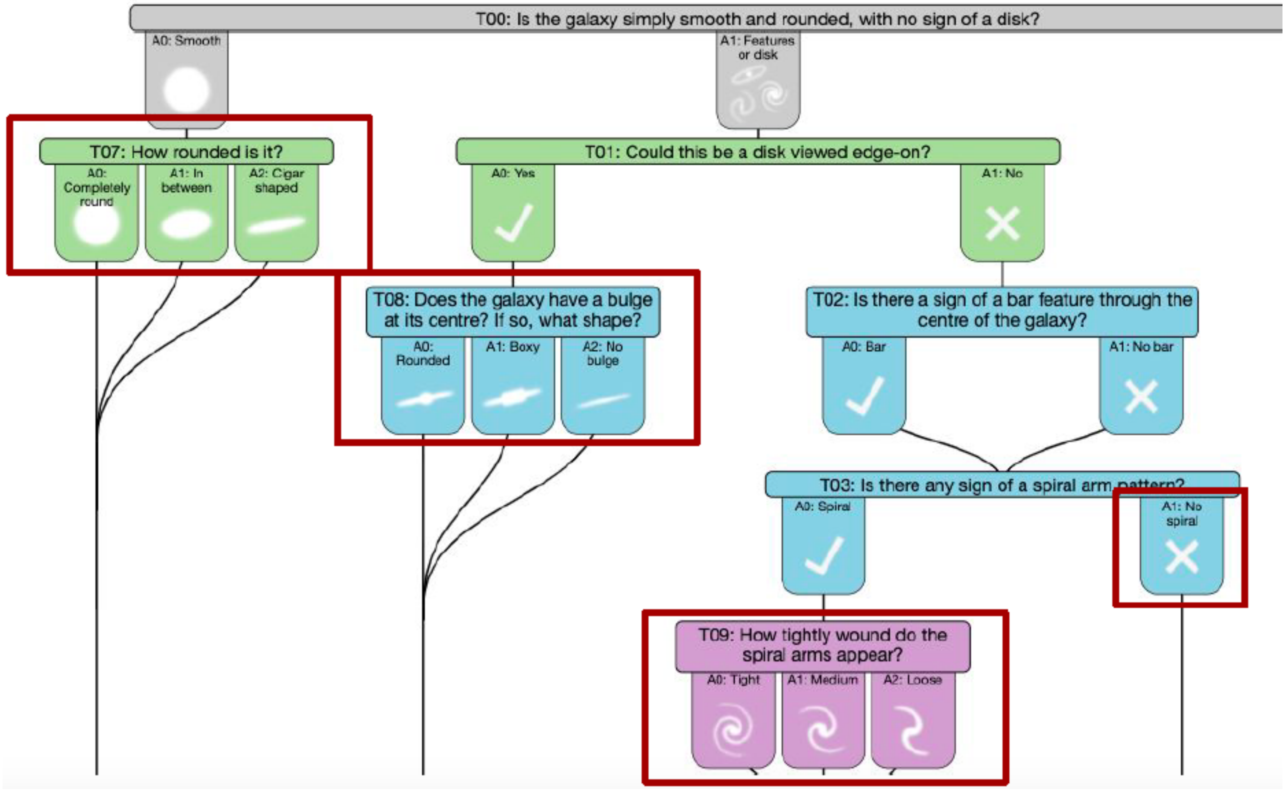
T00: Is the galaxy simply smooth and rounded, with no sign of a disk?
A0: Smooth
A1: Features or disk
T07: How rounded is it?
A0: Completely round
A1: In between
A2: Cigar shaped
T01: Could this be a disk viewed edge-on?
A0: Yes
A1: No
T08: Does the galaxy have a bulge at its centre? If so, what shape?
A0: Rounded
A1: Boxy
A2: No bulge
T02: Is there a sign of a bar feature through the centre of the galaxy?
A0: Bar
A1: No bar
T03: Is there any sign of a spiral arm pattern?
A0: Spiral
A1: No spiral
T09: How tightly wound do the spiral arms appear?
A0: Tight
A1: Medium
A2: Loose

Figure 1. Galaxy classification guide provided by Galaxy Zoo; the project aims to classify galaxies to the ten boxed classes shown above. For the figures in section IV-VIII, I will denote each classification layer as the following:

Layer 1: Smooth / round

Layer 2A: Smooth, completely round / in between / cigar-shaped

Layer 2B: Disk, edge-on / face-on

Layer 3A: Disk, edge-on, rounded bulge / boxy bulge / no bulge

Layer 3B: Disk, face-on, tight spirals / medium spirals / loose spirals

II. Random forest Hyperparameters Tuning

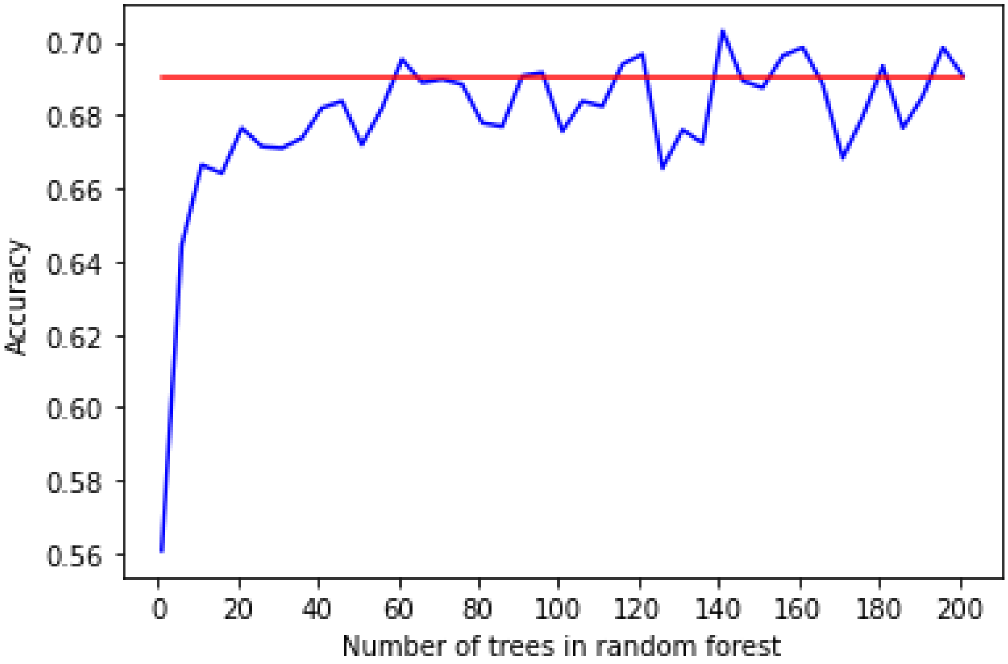

Figure 2. Predictive performance for different number of trees in the random forest

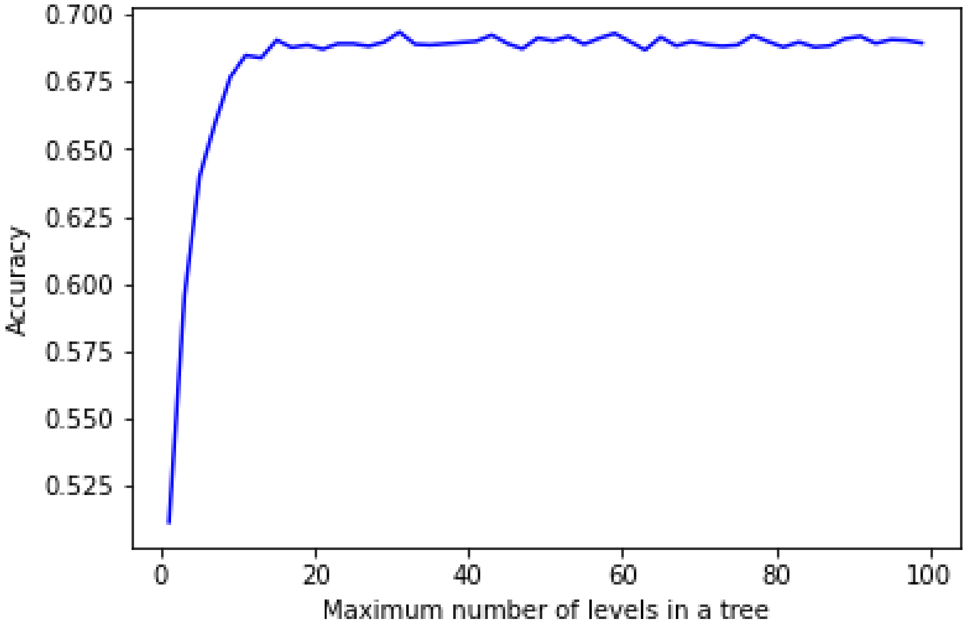

Figure 3. Predictive performance for different number of levels in each tree of the random forest

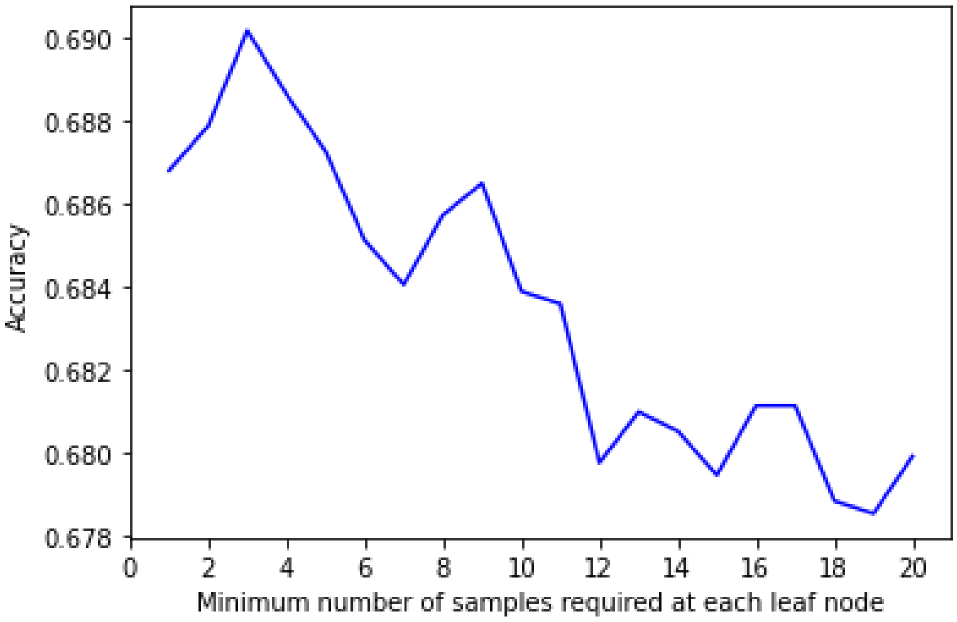

Figure 4. Predictive performance for different minimum number of samples required at each leaf node of every decision tree in the random forest

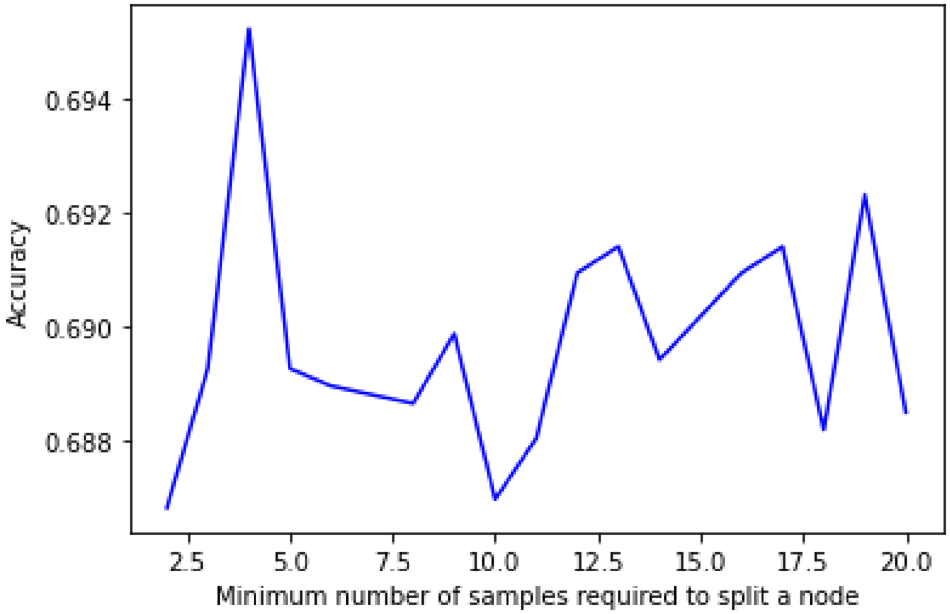

Figure 5. Predictive performance for different minimum number of samples required to split a node of every decision tree in the random forest

III. Overall performance (classifying into all 10 classes)

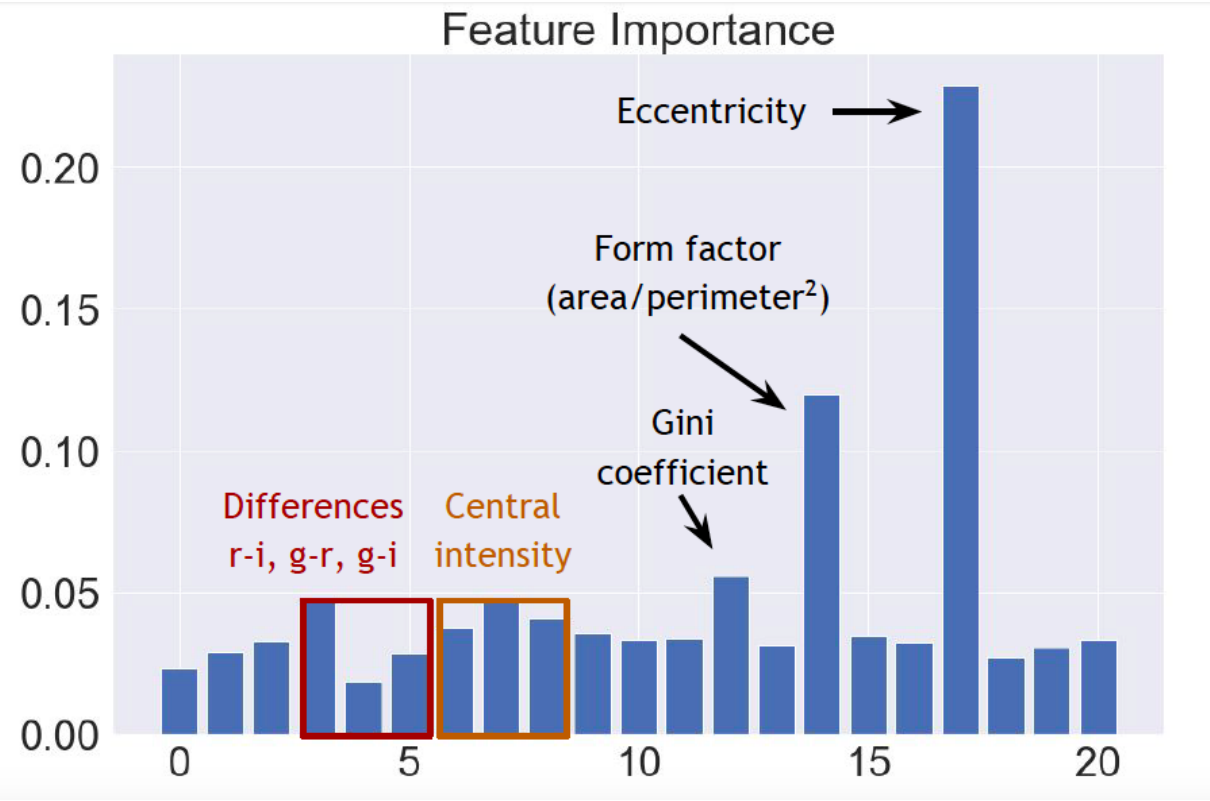

Figure 6. Feature importance plot when classifying galaxies into 10 classes. The most notable features are noted in the plot

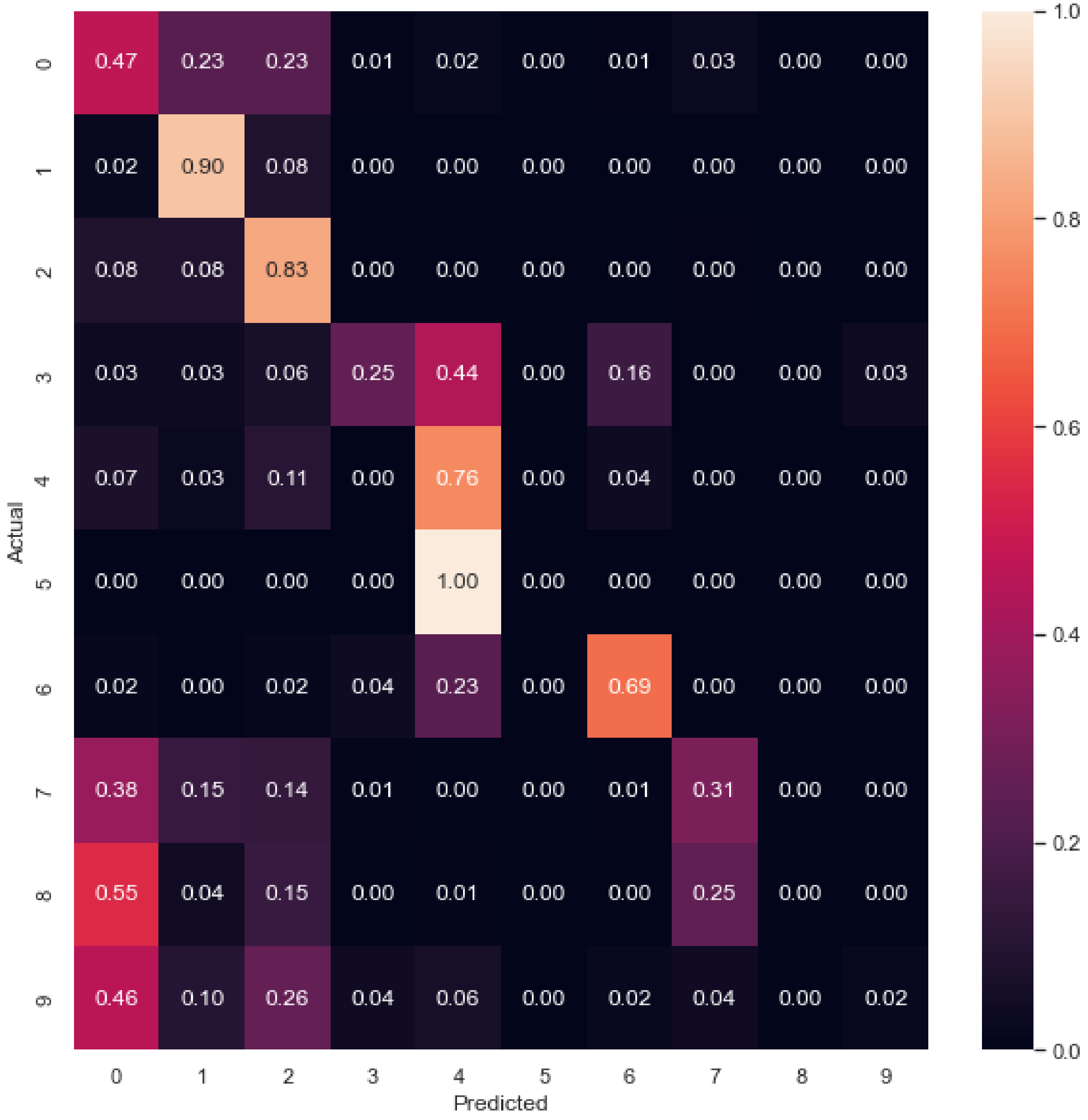

Figure 7. Confusion matrix of the overall classification. Each class number denotes:

0 - Disk, face on, no spiral

1 - Smooth, completely round

2 - Smooth, in-between round

3 - Smooth, cigar shaped

4 - Disk, edge on, rounded bulge

5 - Disk, edge on, boxy bulge

6 - Disk, edge on, no bulge

7 - Disk, face on, tight spiral

8 - Disk, face on, medium spiral

9 - Disk, face on, loose spiral

IV. Classification between disk and smooth galaxies (classification layer 1)

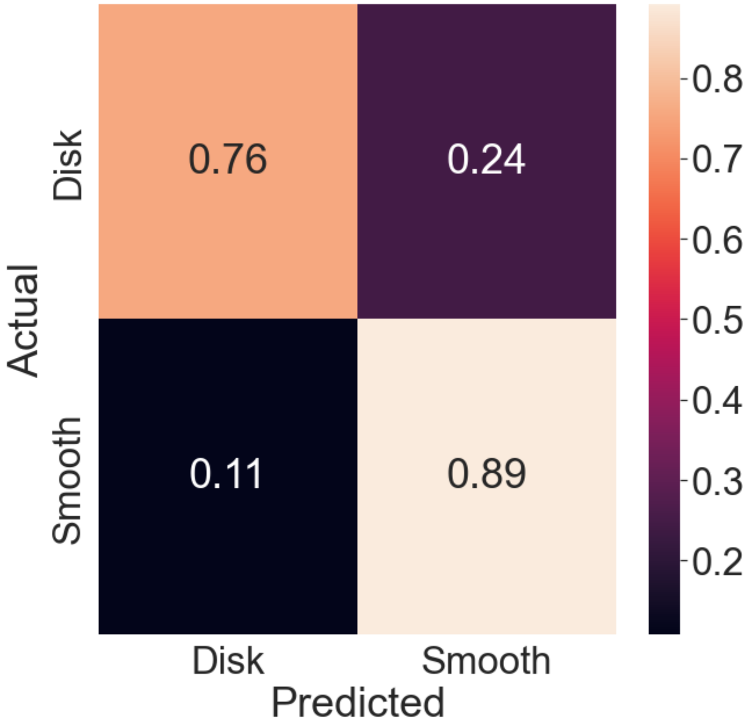

Figure 8. Confusion matrix of classification layer 1

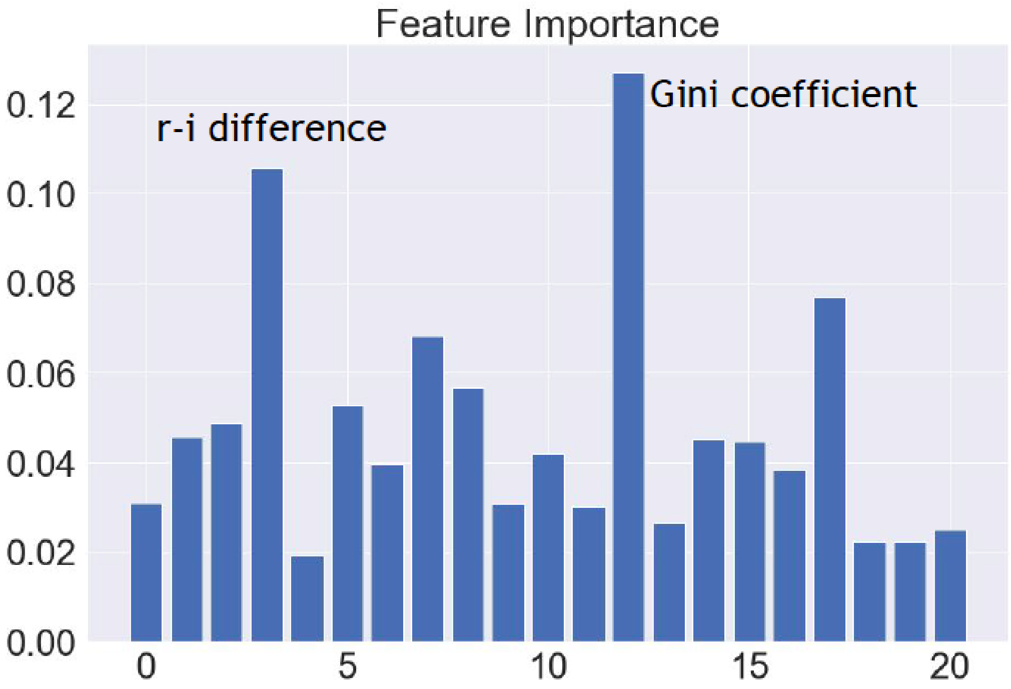

Figure 9. Feature importance of classification layer 1

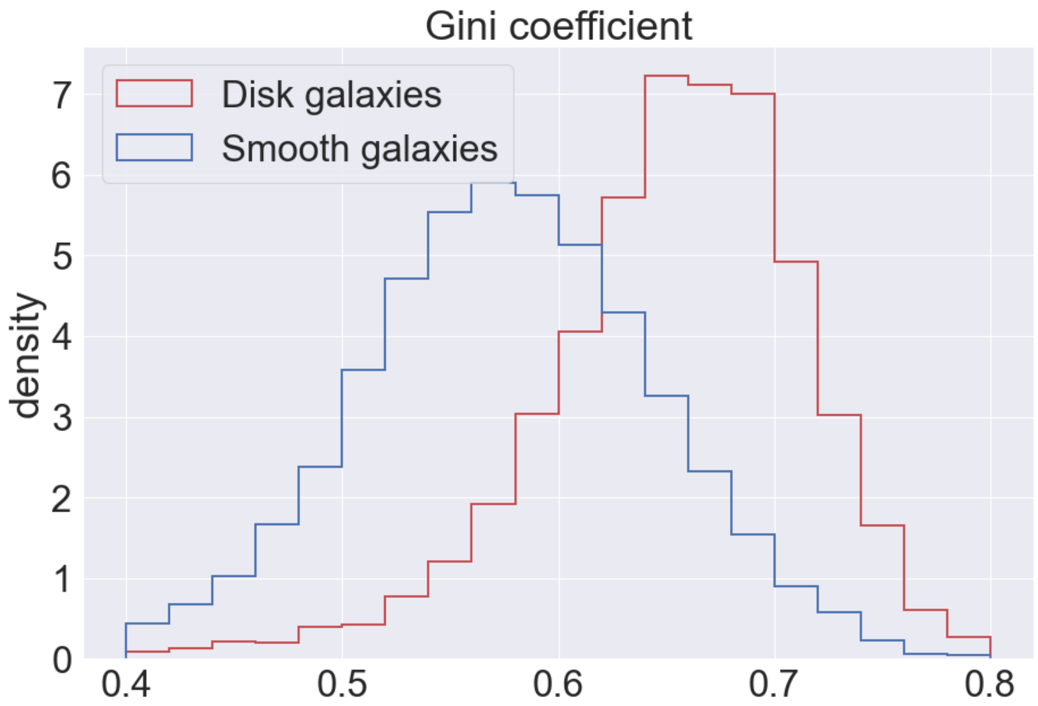

Figure 10. Normalized histograms of the Gini coefficients among disk and smooth galaxies

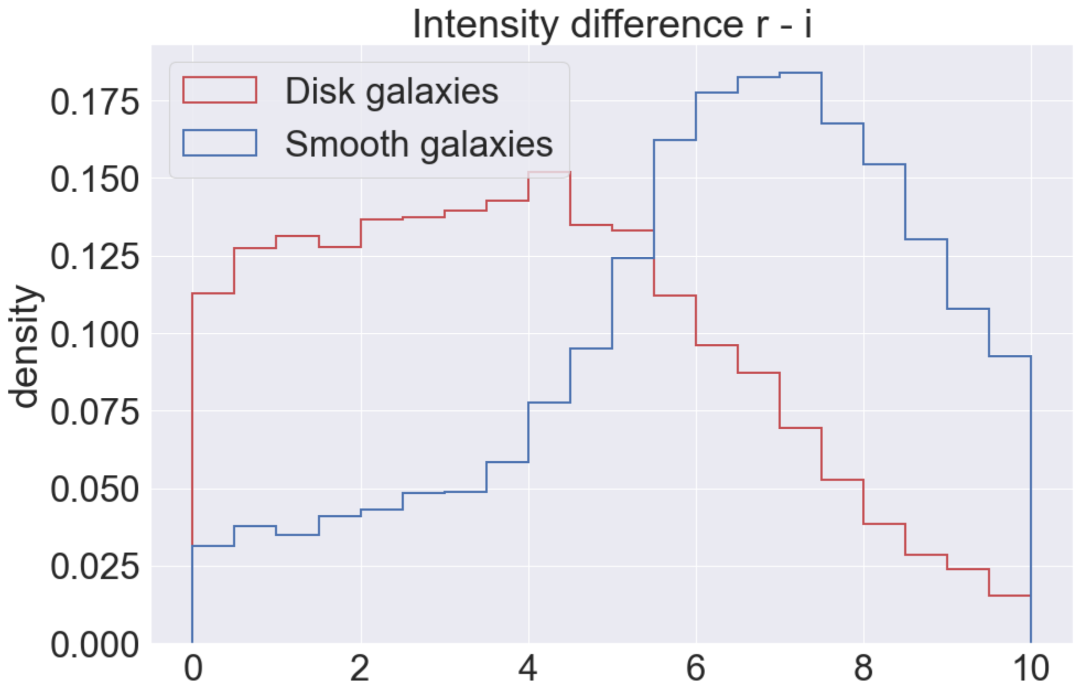

Figure 11. Normalized histograms of the intensity differences between the r and i color bands among disk and smooth galaxies

V. Classification between completely round, in between round, and cigar-shaped galaxies among smooth galaxies (classification layer 2A)

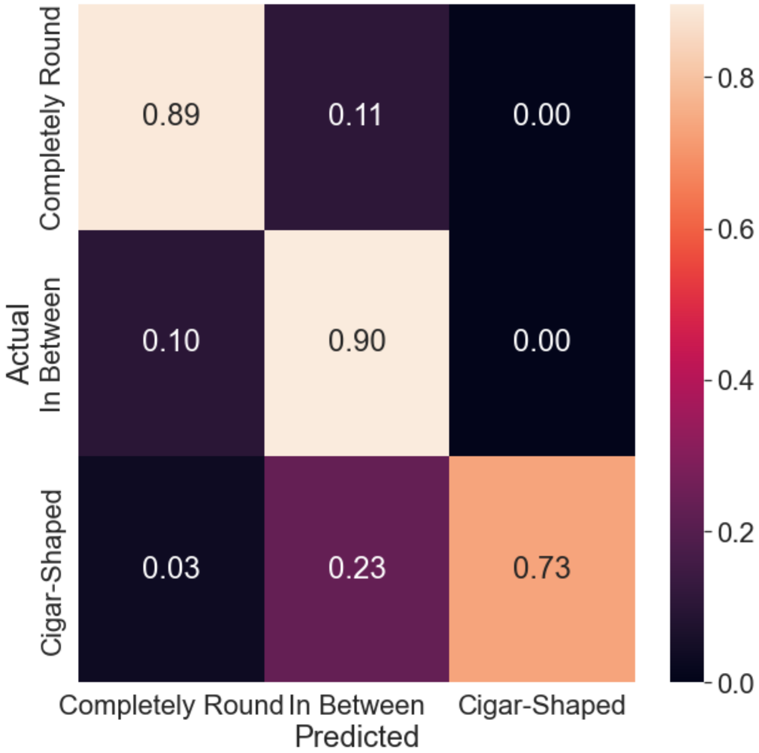

Figure 12. Confusion matrix of classification layer 2A

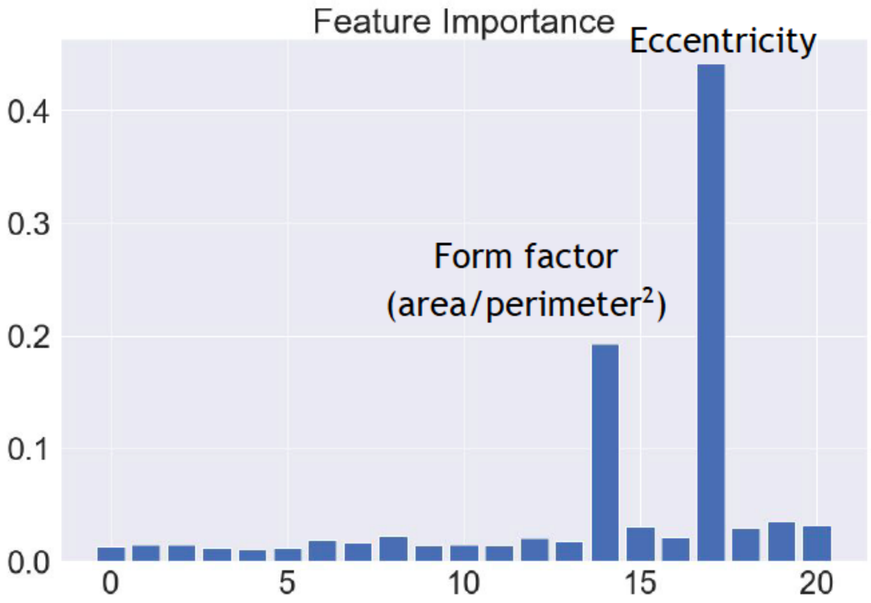

Figure 13. Feature importance of classification layer 2A

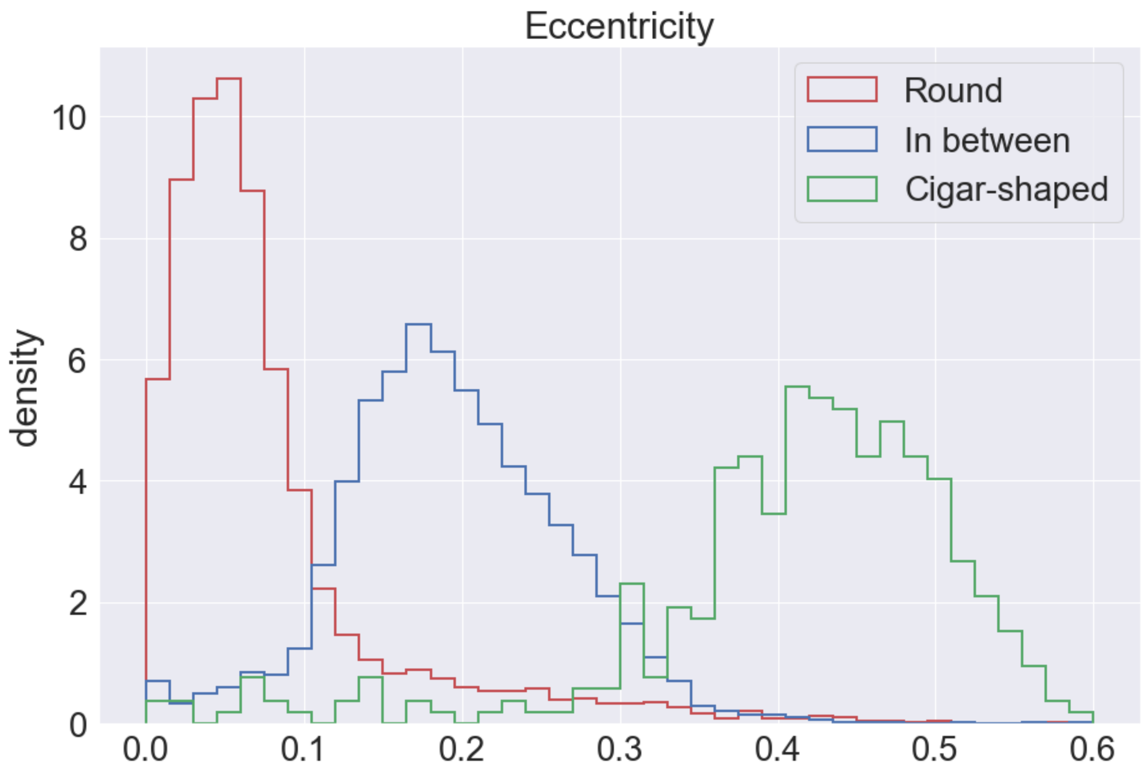

Figure 14. Normalized histograms of the eccentricities among smooth galaxies

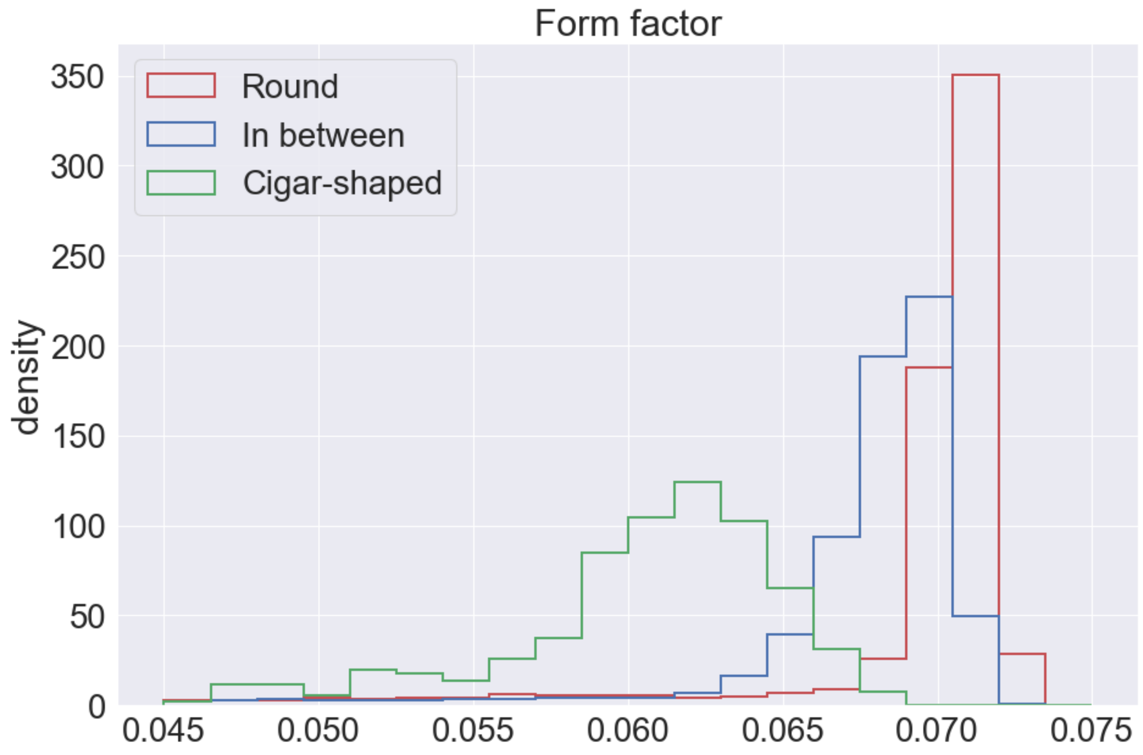

Figure 15. Normalized histograms of the form factors among smooth galaxies

VI. Classification between edge-on and face-on galaxies among disk galaxies (classification layer 2B)

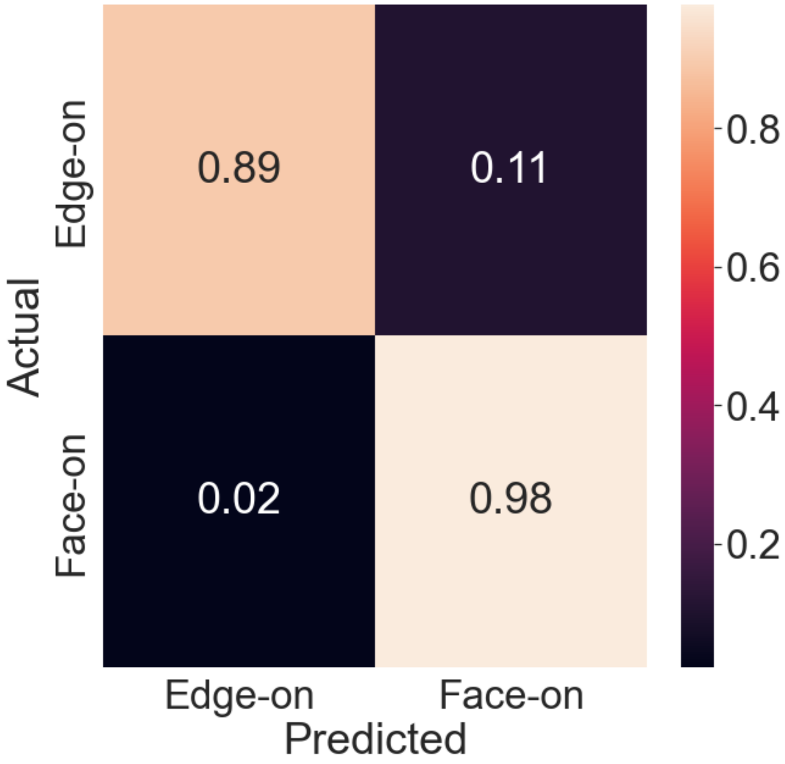

Figure 16. Confusion matrix of classification layer 2B

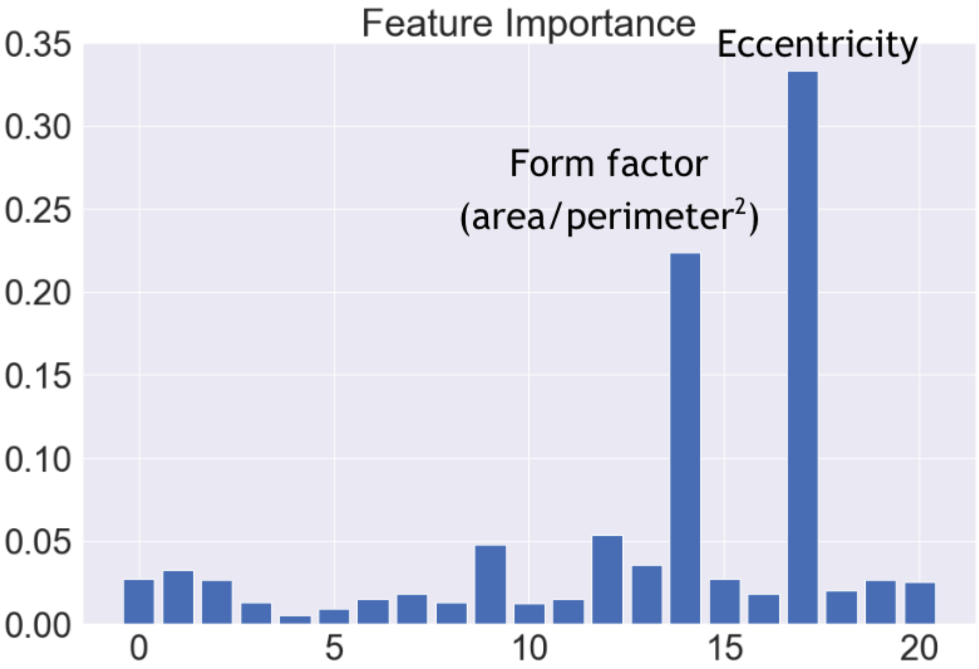

Figure 17. Feature importance of classification layer 2B

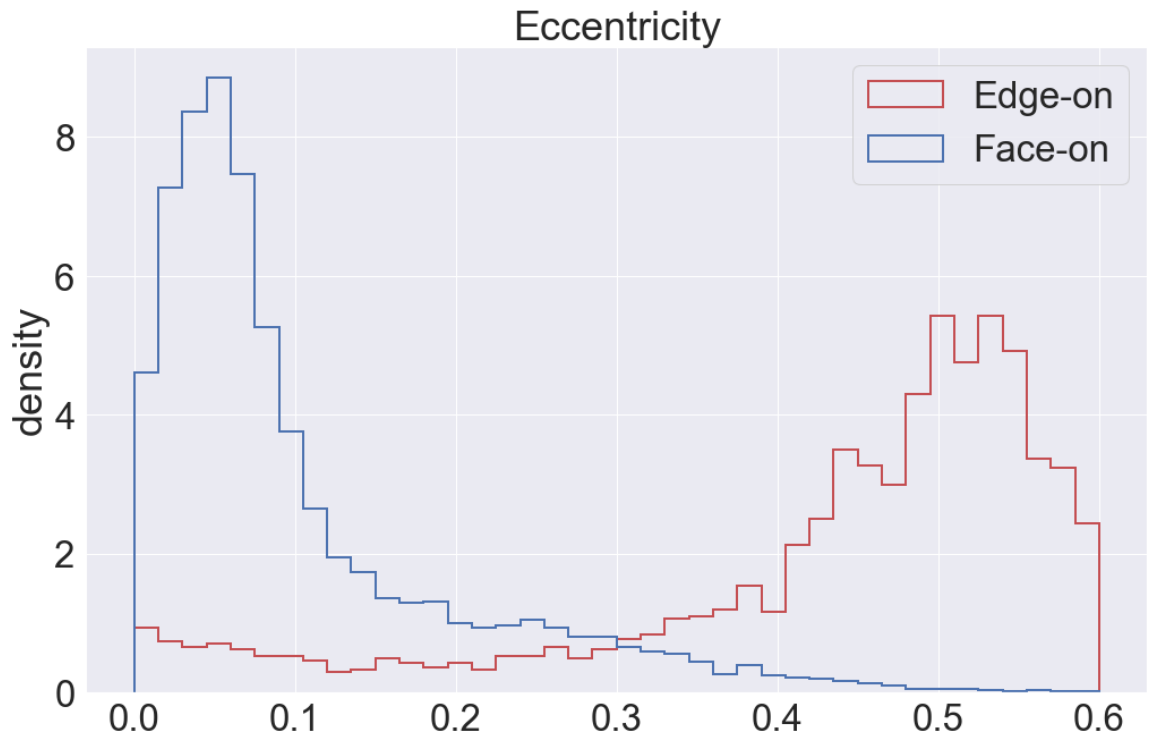

Figure 18. Normalized histograms of the eccentricities among disk galaxies

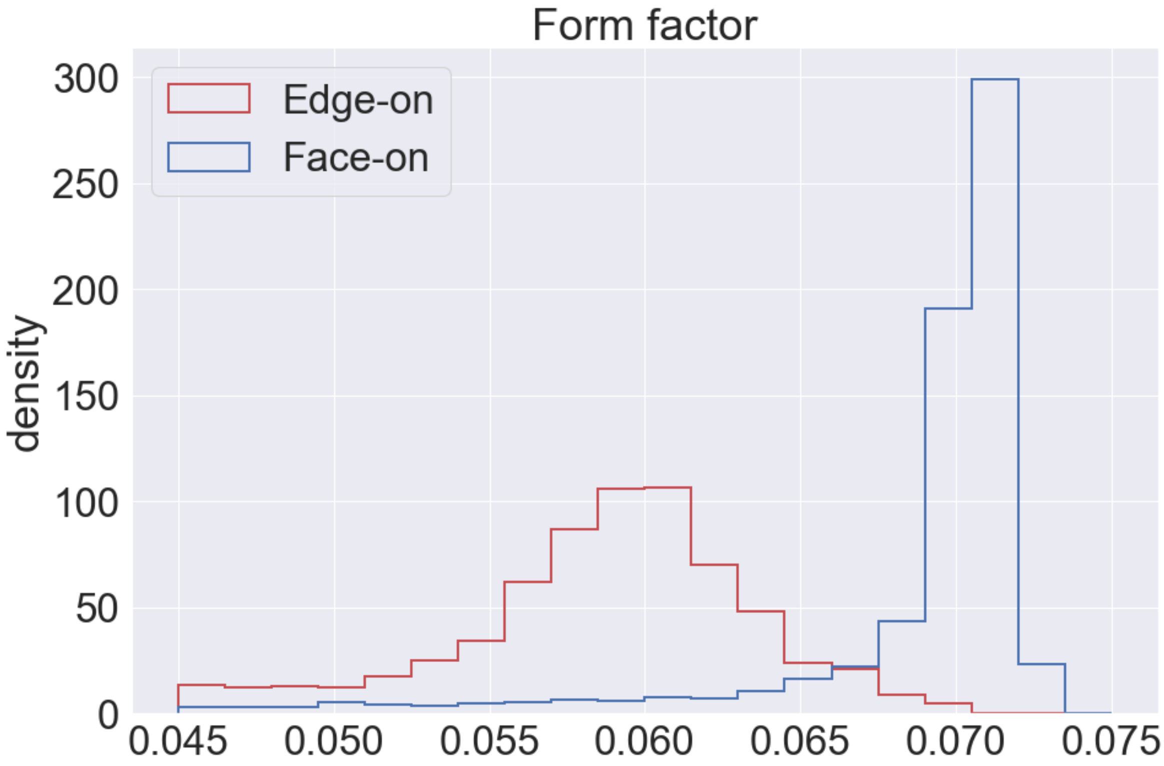

Figure 19. Normalized histograms of the form factors among disk galaxies

VII. Classification between round bulge, boxy bulge, and no bulge among disk and edge-on galaxies (classification layer 3A)

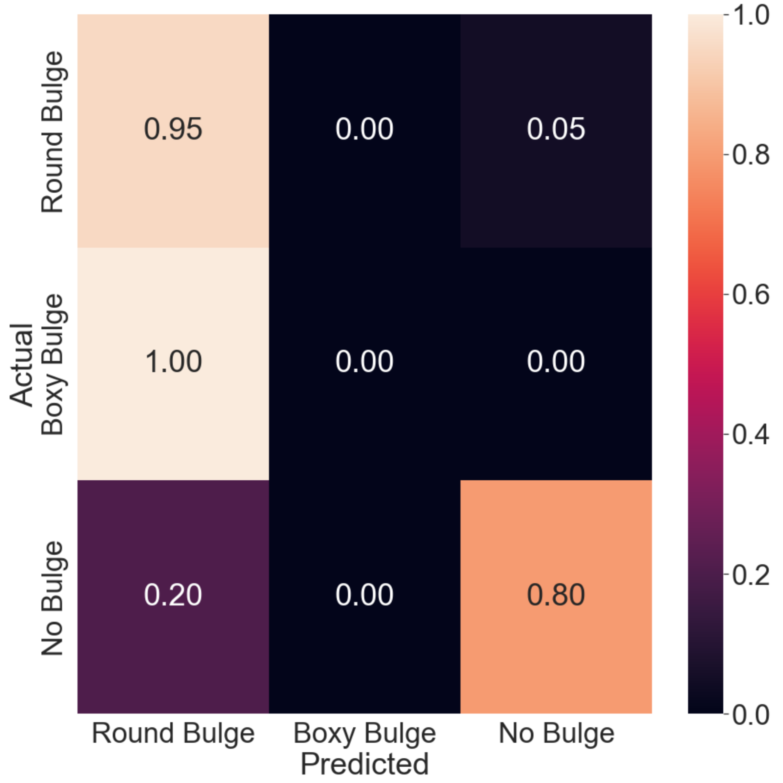

Figure 20. Confusion matrix of classification layer 3A

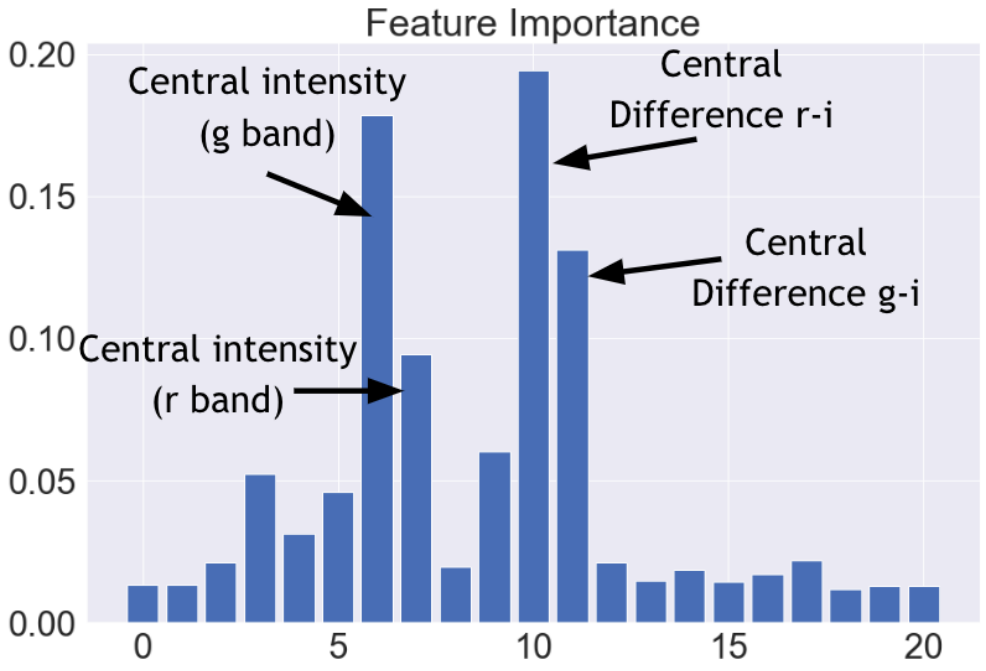

Figure 21. Feature importance of classification layer 3A

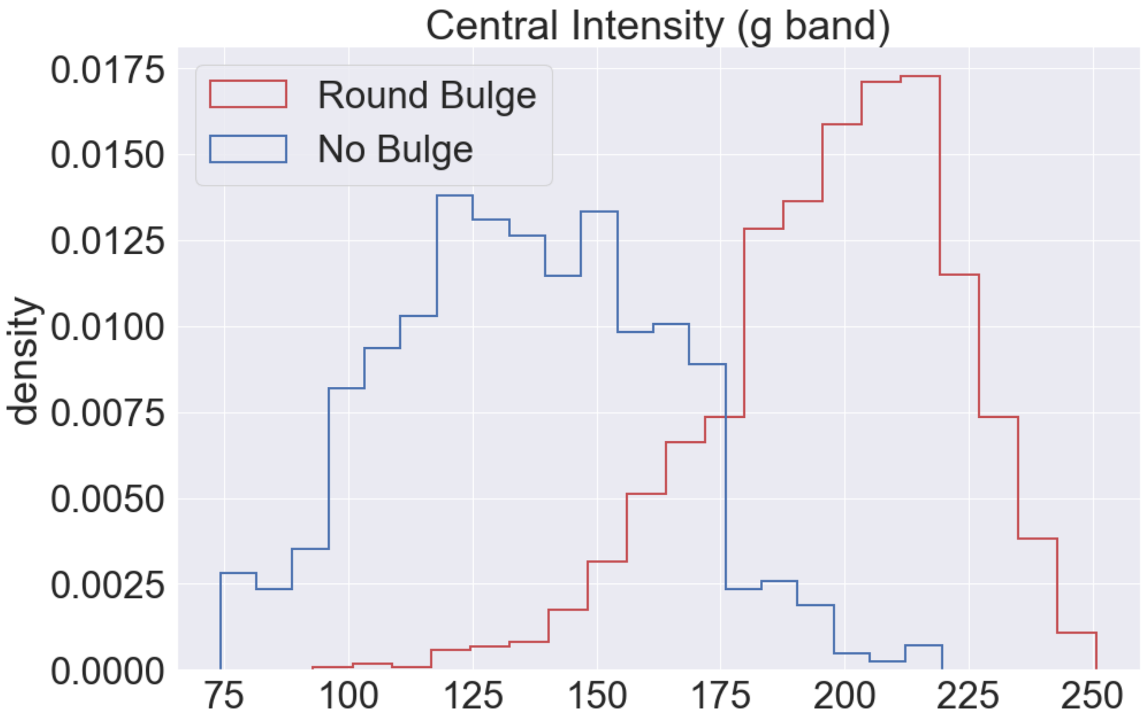

Figure 22. Normalized histograms of the central g-band intensities among disk, edge-on galaxies

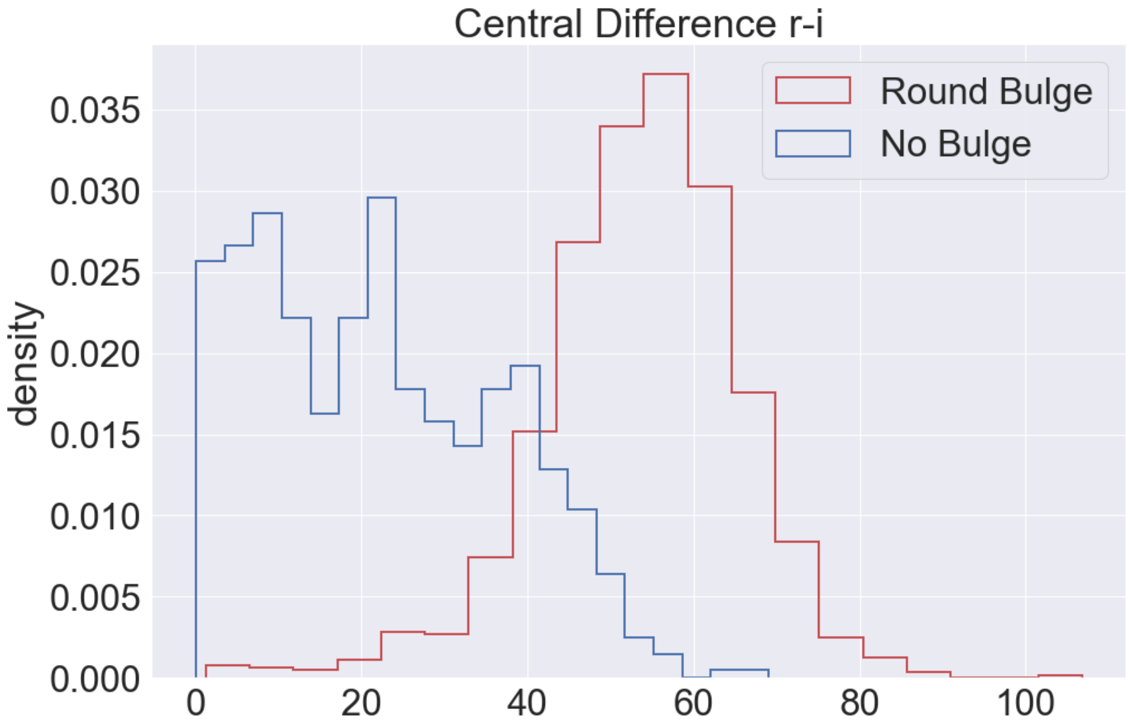

Figure 23. Normalized histograms of the central intensity differences between the r and i color bands among disk, edge-on galaxies

VIII. Classification between no spiral, tight spiral, medium spiral, and loose spiral among disk and face-on galaxies (classification layer 3B)

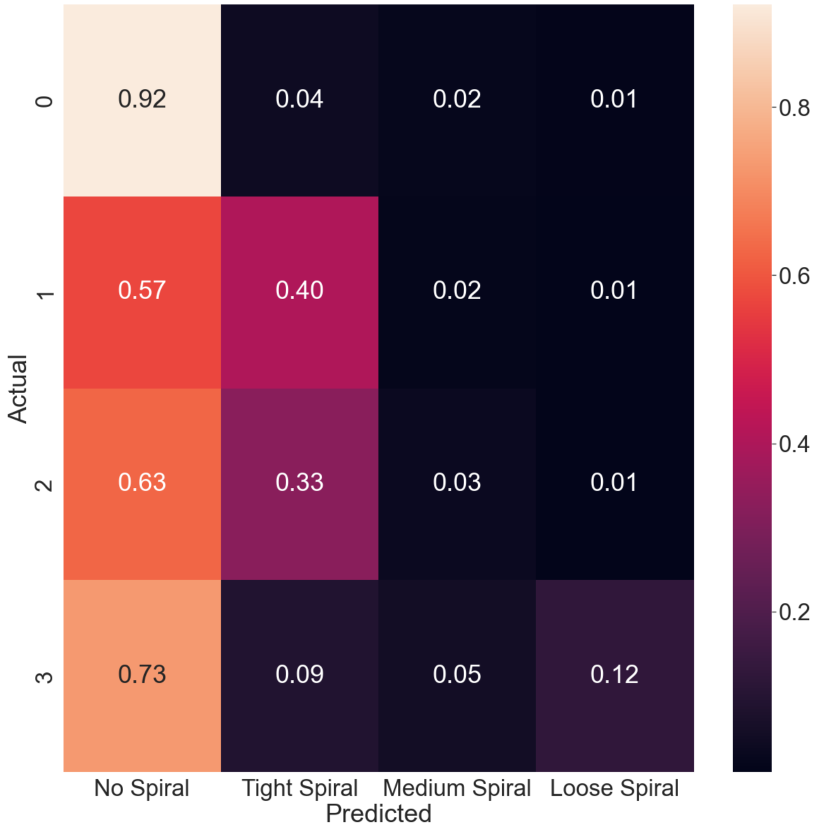

Figure 24. Confusion matrix of classification layer 3B

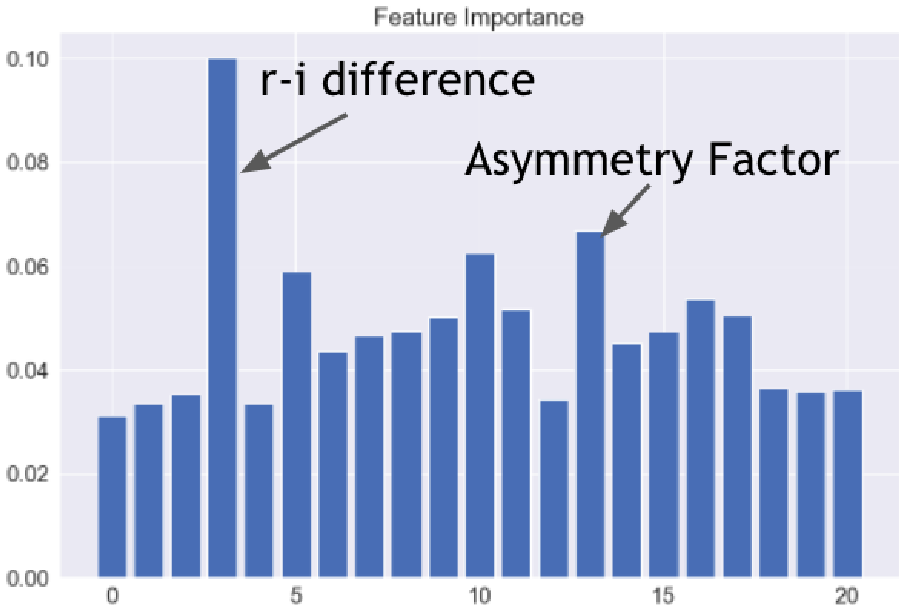

Figure 25. Feature importance of classification layer 3B

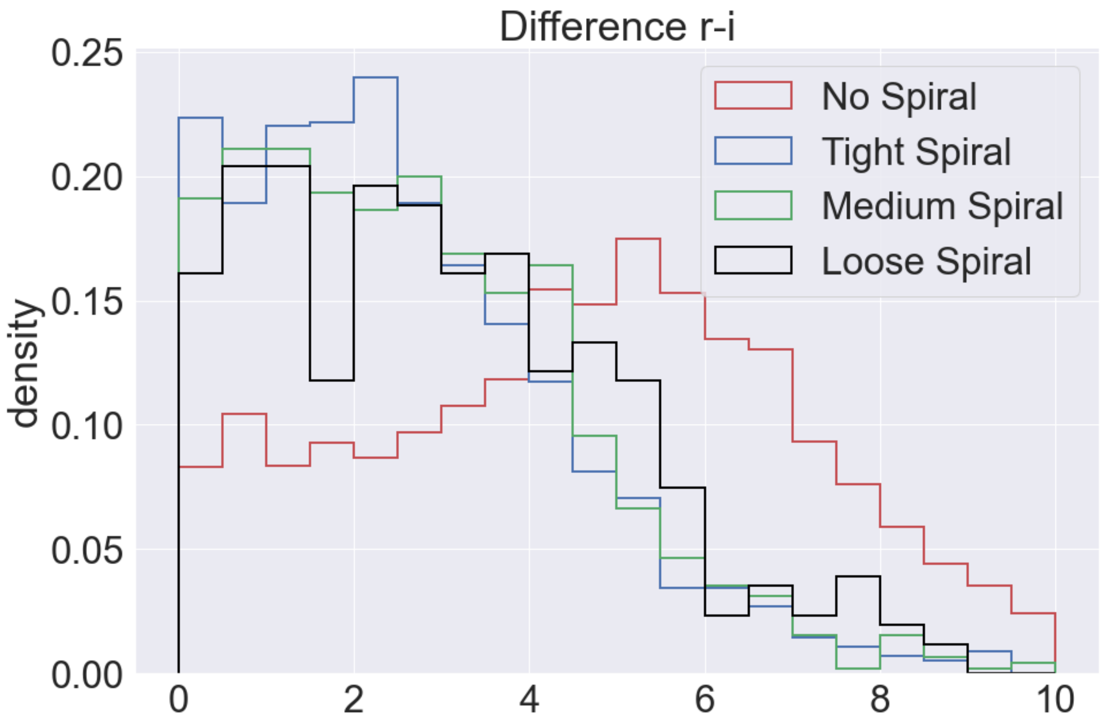

Figure 26. Normalized histograms of the intensity differences between the r and i color bands among disk, face-on galaxies

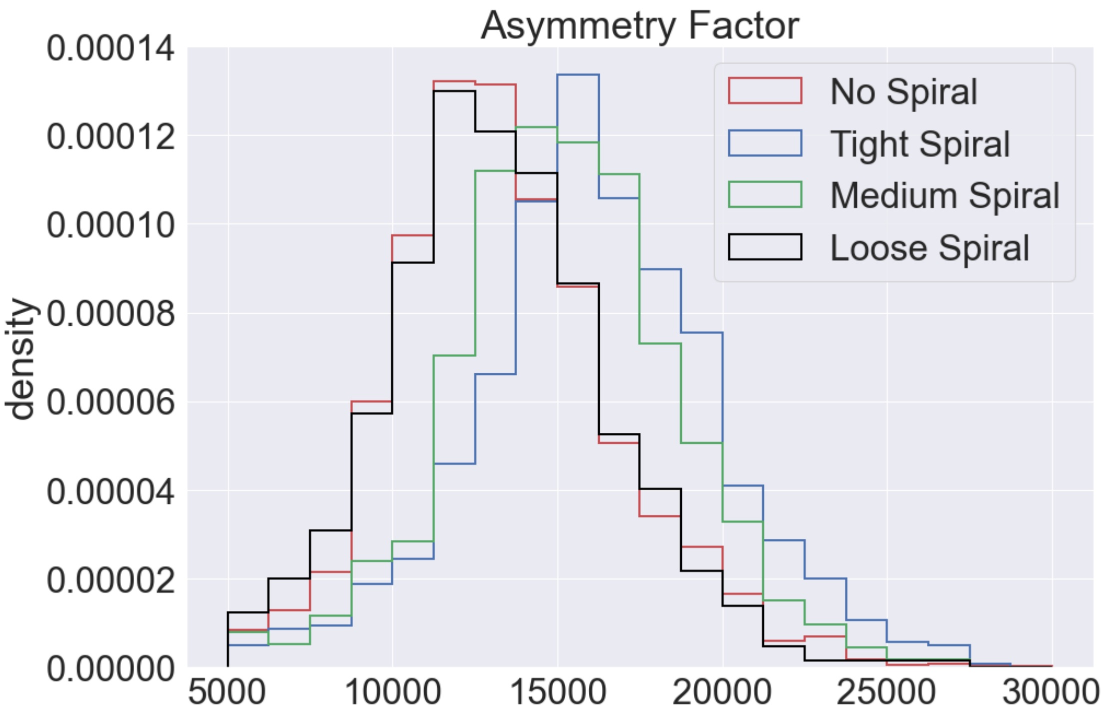

Figure 27. Normalized histograms of the asymmetry factors among disk, edge-on galaxies