# Galaxy Zoo 2: Basic Morphology Classification and Analysis using Deep Learning

Tobin
Department of Computer Science
May 2025

Contents

# 1 Introduction

Galaxy morphological classification is a way that scientists can divide the shapes and forms of galaxies into families. These families are based upon several factors, including their density, angle, formation characteristics, and a few other factors. One way of going about this categorization is with the Hubble Sequence, using shorthand to classify galaxies. For example, a galaxy with the designation SBa is a spiral-barred galaxy with tightly wound and smooth arms. The Galaxy Zoo 2 (gz2) dataset is a large (~300k) collection of RGB images of galaxies that have been classified by hand by volunteer "citizen" scientists. This was done using a decision tree.

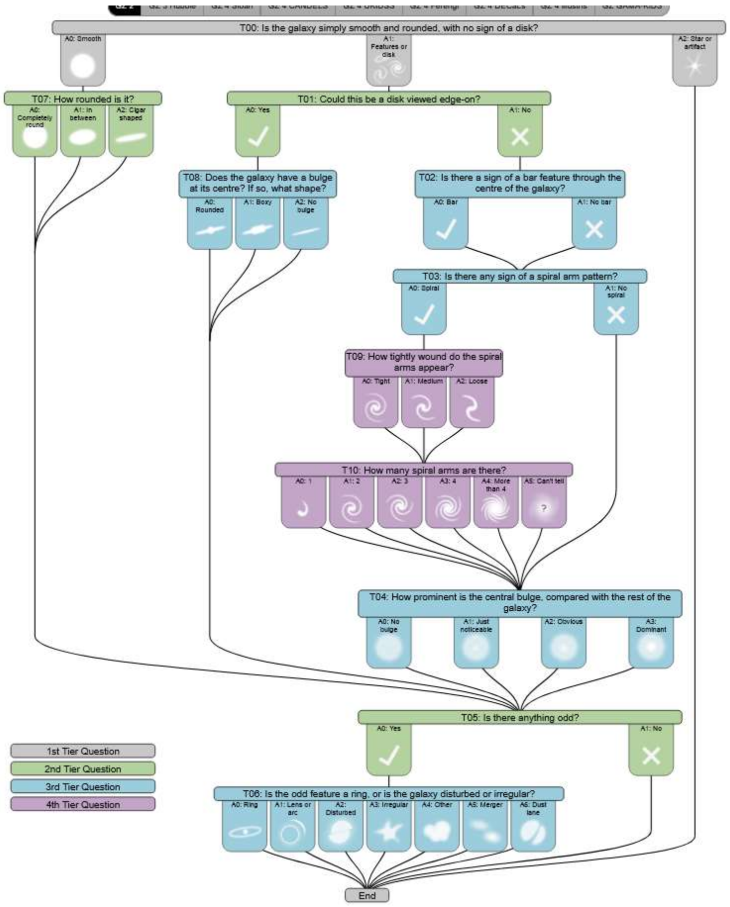
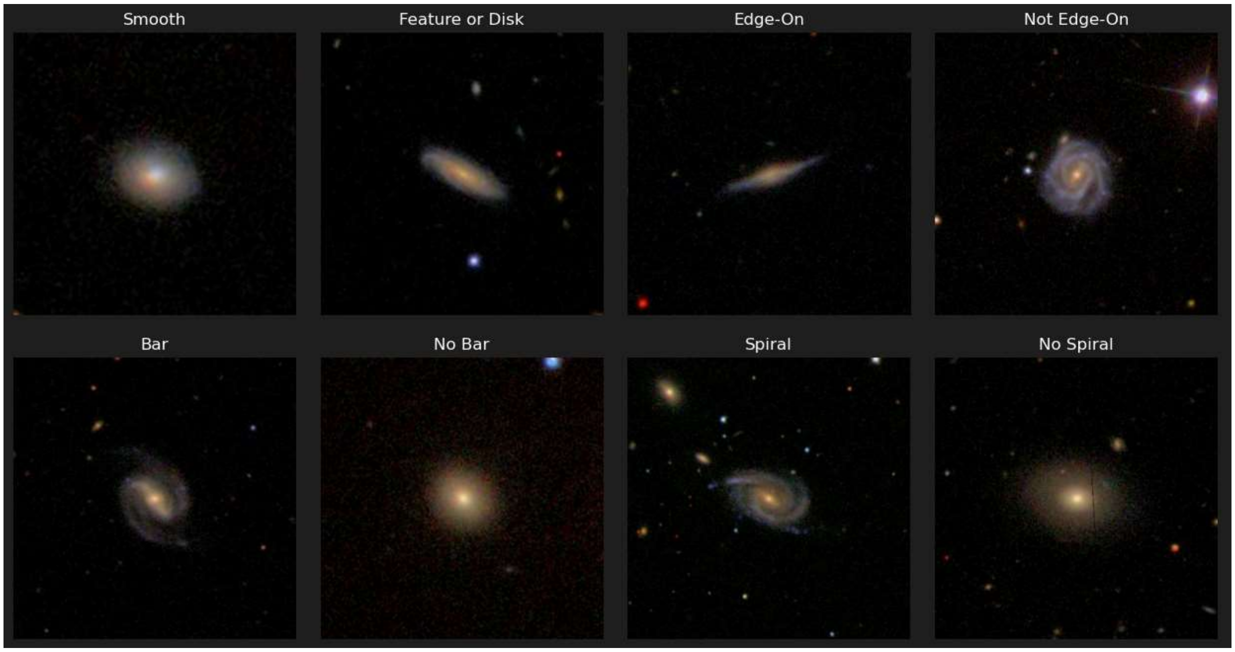

T00: Is the galaxy simply smooth and rounded, with no sign of a disk?
A0: Smooth
A1: Features or disk
A2: Star or artifact
T07: How rounded is it?
A0: Completely round
A1: In between
A2: Cigar shaped
T01: Could this be a disk viewed edge-on?
A0: Yes
A1: No
T08: Does the galaxy have a bulge at its centre? 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
T10: How many spiral arms are there?
A0: 1
A1: 2
A2: 3
A3: 4
A4: More than 4
A5: Can't tell
?
T04: How prominent is the central bulge, compared with the rest of the galaxy?
A0: No bulge
A1: Just noticeable
A2: Obvious
A3: Dominant
T05: Is there anything odd?
A0: Yes
A1: No
1st Tier Question
2nd Tier Question
3rd Tier Question
4th Tier Question
T06: Is the odd feature a ring, or is the galaxy disturbed or irregular?
A0: Ring
A1: Lens or arc
A2: Disturbed
A3: Irregular
A4: Other
A5: Merger
A6: Dust lane
End

Figure 1

Individuals would classify each galaxy they were shown, following this tree. Some questions were simple, yes or no, and others involved a bit more in-depth classification, such as the oddities. These answers were then pooled together for each image to get an overall classification.

For this project, it will focus on the wider classifications of differentiating between smooth or featured, edge-on or off, barred or not barred, and spiraled or not spiraled galaxies.



# 2 Data Preprocessing

## 2.1 Pruning the Data

Due to the nature of the data classification of gz2, the determination of an image is based on the majority vote for each question asked. The data was structured to have a count of votes for each label, including each no and yes.

For example, a sample may have the labels:

t01_smooth_or_features_a01_smooth_count = 41

t01_smooth_or_features_a02_features_or_disk_count = 11

Since these are a conflicting datapoint for a binary question, gz2 includes a debiased fraction of the votes for each feature. The galaxy in question is then classified based upon this weighted fraction (> 50%).

t01_smooth_or_features_a02_features_or_disk_debiased = 0.788..

t01_smooth_or_features_a03_smooth_debiased = 0.211…

This majority vote means there are large amounts of the data points that are deemed uncertain. Therefore, all decisions were limited to a Count >= 10 and Fraction >= 0.80 for their respective labels.

Using samples with N values less than the given resulted in training "sticking" at any accuracy found after the initialization of weights, and increasing the epoch amount showed no improvements, even up to 400 epochs.

## 2.2 Data balancing

For each classification, there were different numbers of entries for each label and class. Table 1 shows the total of the label and each class.

| Label | Total Entries | Per Class Amount |
|---|---|---|
| Shape | 154,664 | Smooth: 122,158<br>Featured/Disk: 32,506 |
| Edge Facing | 67,281 | On: 10,038<br>Off: 57,243 |
| Barred | 31,594 | Barred: 5,457<br>No Bar: 26,137 |
| Spiraled | 43,171 | Spiral: 35,858<br>No Spiral: 7,316 |

Table 1

To account for the imbalances of classes, the total value of each class was capped at the minimum amount of the classes between corresponding values.

## 2.3 Augmentation

Upon constructing each part of the dataset, train, validate, and test, augmentations were applied to ensure the models learn general cases of each category. Random flipping of both horizontal and vertical directions as well as random rotation of any degree [0-360].

### 2.3.1 Forgone Augmentations

No image stretching was implemented, as galaxies naturally can be stretched, so image stretching may cause more overfitting. No greyscale conversion was used as well, because the eventual goal outside of this class is to create a model that can identify images taken with my telescope. Greyscale could be used for this case; however, due to the number of galaxies in the dataset, I figured the models would not overfit based on colors, even though some attributes are more common with certain color galaxies.

# 3 Model

Two different models were used for the testing on the dataset: a custom CNN-based architecture, as well as ResNet50 feature extraction passed into a random forest model.

## 3.1 Model Architecture

The CNN model has three blocks of two convolutional layers, followed by a max pooling layer. After all three layers, there is a layer of average pooling followed by two dense layers, with the final condensing down to two neurons with an activation function of sigmoid. The optimizer for this model is Adam, and the loss is sparse_categorical_crossentropy.

The random forest-based model passed all the images into ResNet50, extracted features, and passed them into TensorFlow's Random Forest Classifier.

## 3.2 Hyperparameter Tuning

For the CNN model itself, differing amounts of convolutional layers, complexity, and filter sizes were used to determine which is best. Unfortunately, due to the lack of time there was less experimentation with the random forest classifier, and the only tested hyperparameter was testing the number of estimators between 50 and 100.

### 3.2.1 CNN Hyperparameters

The number of convolutional networks with this model was tested thoroughly. The resulting amount was finalized at six, but nine and even twelve were also attempted. The only meaningful difference this was found to make was a reduction in Google Collab Pro compute units.

The order was changed, as the model increases in complexity, the number of filters grows 32-64-128. The order of these was changed and played with (128-64-32); however, like above, it just used compute units with poor results.

The number of epochs for this model is changed, however, ranging from four to ten, depending on which category it was used to classify. This will be expanded upon later, however, it did not take many epochs to train each category.

The optimizer was Adam, and the loss equation was sparse categorical cross-entropy. Different optimizers, such as SDG and Adagrad, however, Adam won overall and was not replaced after experimentation. The loss equation of Sparse Categorical Crossentropy was used even though it is more suited for multiclass labeling rather than binary crossentropy, as I would like to expand this model further in the future. However, this being stated, there was no noticeable difference in the two, so sparse was kept.

The batch size of the dataset was also messed with, ending with a size of 64. This did not seem to change the overall accuracy of the model when < 64, but did increase runtime heavily.

### 3.2.2 Random Forest Hyperparameters

Unfortunately, due to my lack of time, I was unable to test as much as I would've preferred to for my Random Forest hyperparameters. The number of estimators used was one hundred; lessening this did reduce accuracy, but one hundred did not take too long to run on each of my datasets.

### 3.2.3 Training, Validation, and Test Set Split

The data was split into 70-15-15. Due to the size of the dataset alongside augmentation, this was the only split tested.

# 4 Experiments

Four binary characteristics were used for these experiments. Allowing the CNN and RF models to categorize based only on the two. Each binary classification was separate from the others as there was trouble balancing the dataset for all classes.

## 4.1 Shape

Classifying the shape of a galaxy was the most successful in both my CNN and RF models. The number of epochs used was only four for this classification. More epochs saw a greater accuracy, but gains were diminishing, and due to the size of this first dataset, it took too long to reasonably continue. Figure 2 shows the accuracy gains of the convolutional model, ending with a total validation accuracy of 95.05%. Although the graph does not show it, the cap of accuracy was reached here with only small gains of around half a percent found after following epochs. The epochs were truncated just for simplicity's sake of runtime.

A test was run on the created test set with an accuracy of 94.84%, showcasing the success of smooth classifications.

The Random Forest classifier also saw a relatively high overall accuracy of 88%, however had a bad recall score for the Featured/Disk classifier at 55%. I believe this is due to a mismatch in the balancing, however, time has prevented further testing.

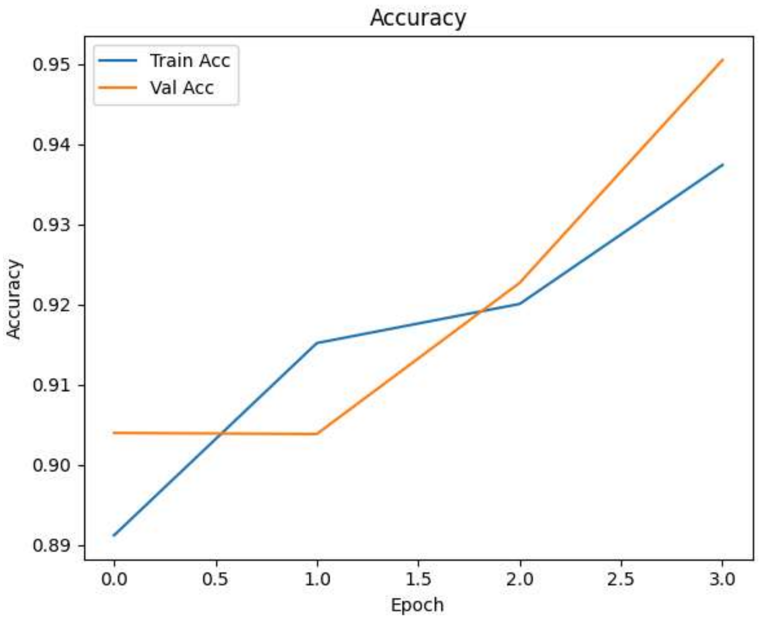


Figure 2

## 4.2 Edge

The CNN training of edge-on or edge-off galaxies also had good results, reaching a 94.65% accuracy after six epochs. Just like the shape classifier, only six epochs were used as returns diminished and runtime increased. Figure 3 partially shows the diminishing returns that were seen with more epochs. It remains to be seen how close to 100% accurate this could be, however, training was stopped just for the preservation of my compute units.

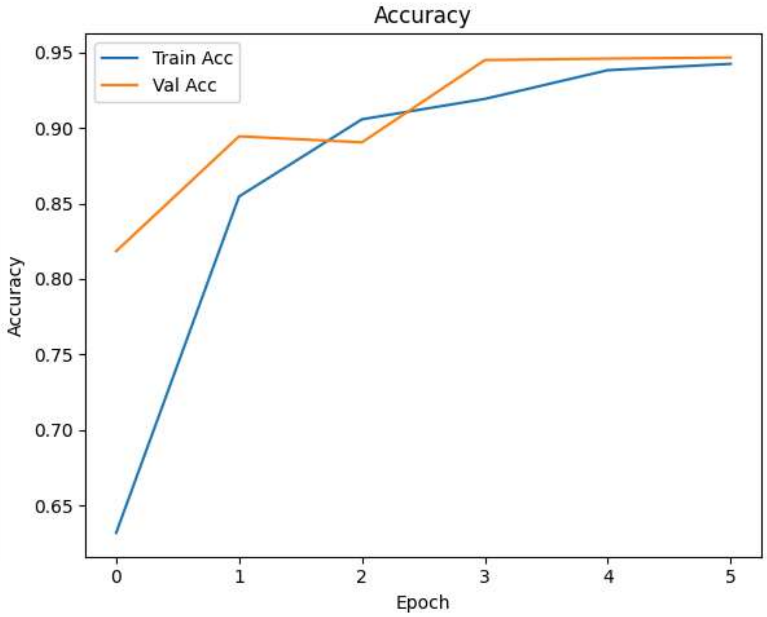


Figure 3

Random Forest also saw similar results for this, with an overall accuracy of  88% as well.

## 4.3 Barred

Coming to the barred category, I saw a substantial drop in the ability of my model. I did a large amount of testing, involving changing all hyperparameters, increasing epochs, and increasing the counts and debiased fraction of the data. No matter what I did, I could barely get better than the random initialization of my models. I would've liked to have continued testing, but due to the time it took to get the initial models working, I am not able to further improve this under the time constraints for this assignment. The maximum accuracy of my model was 65.12%, marginally better than flipping a coin. Figure 4 shows the truncated graphs of the loss and accuracy of this classification. I did attempt to run this with a large number of epochs (400), and there was no improvement. Figure 5 shows the loss and accuracy of the 400-epoch model that I did just for fun. It overfitted during this, but still, the accuracy was not good. Dropout was added with the 400 epochs, but overfitting still occurred, although it didn't really matter.

Random Forest saw a very marginal improvement with a precision of 66% and 68% for edge on and off, respectively. The overall accuracy was 67%, which is slightly better but not enough to justify calling this complete. I would like to come back to this in the future if I do have the time.

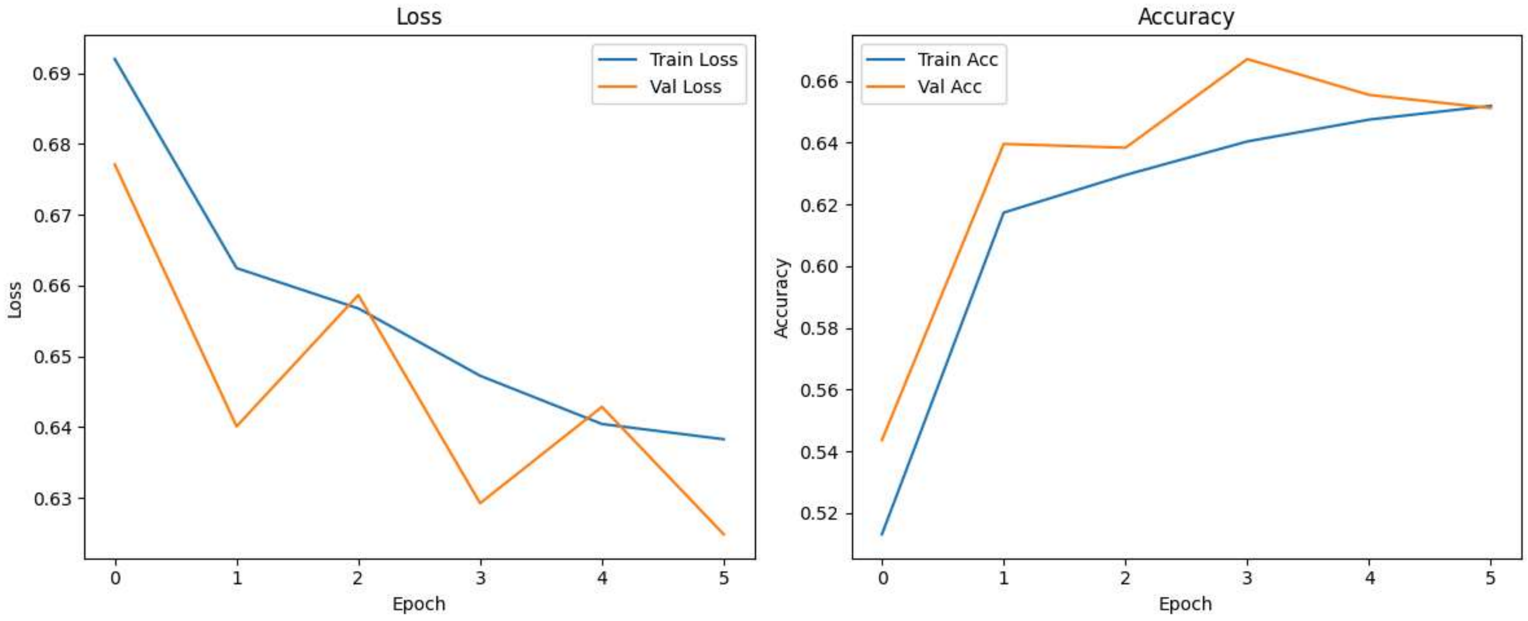


Figure 4

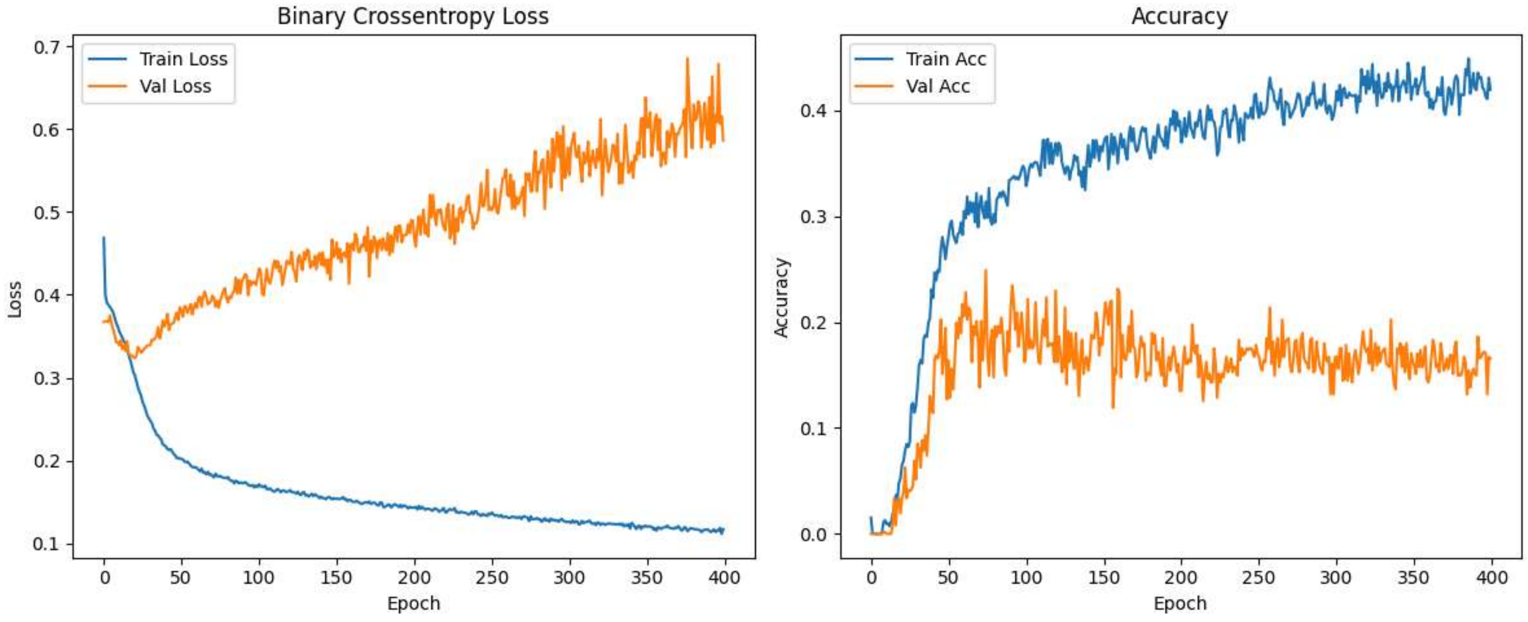


Figure 5

## 4.4 Spiraled

For spiraled, the results from both the CNN model showed an improvement compared to the barred classification, whereas the Random Forest model differed from the CNN and showed similar performance to the barred model. With 10 epochs, a validation accuracy of 74.52% was achieved. This is still not perfect, and several reasons could be relevant for preventing better classification. I was surprised by this fact, as the human eye spiraled galaxies are much easier to detect with the Gz2 dataset compared to barred galaxies. Figure 7 showcases my results from this training. I was skeptical of this result and figured it must come from a way I was interpreting the dataset. To test this, I created a different model using the original Galaxy Zoo dataset. Gz2 is interpreted using a weighted fraction of votes to classify each galaxy, whereas Gz1 is completely determined on "SPIRAL, ELLIPTICAL, UNCERTAIN." Only galaxies that are certainly spiral are classified as such. With this in mind, I created my model only comparing Spiral and Elliptical galaxies and had great results. With 10 epochs, I achieved an accuracy of ~95%, as seen in Figure 8. Of course, this is a different dataset, however, it removed ambiguity from the training. I am unsure if this is what fully caused my problem, but as I am low on time, I decided to try to figure this out after writing this report. As a side note, the average fraction of votes for a spiral galaxy was ~84%, so the CNN model is not far off the split of the citizen scientist votes.

As for the Random Forest model, it achieved precisions of 63% and 68% for sprialed and non-sprialed, respectively, with an overall accuracy worse than that of my CNN model.

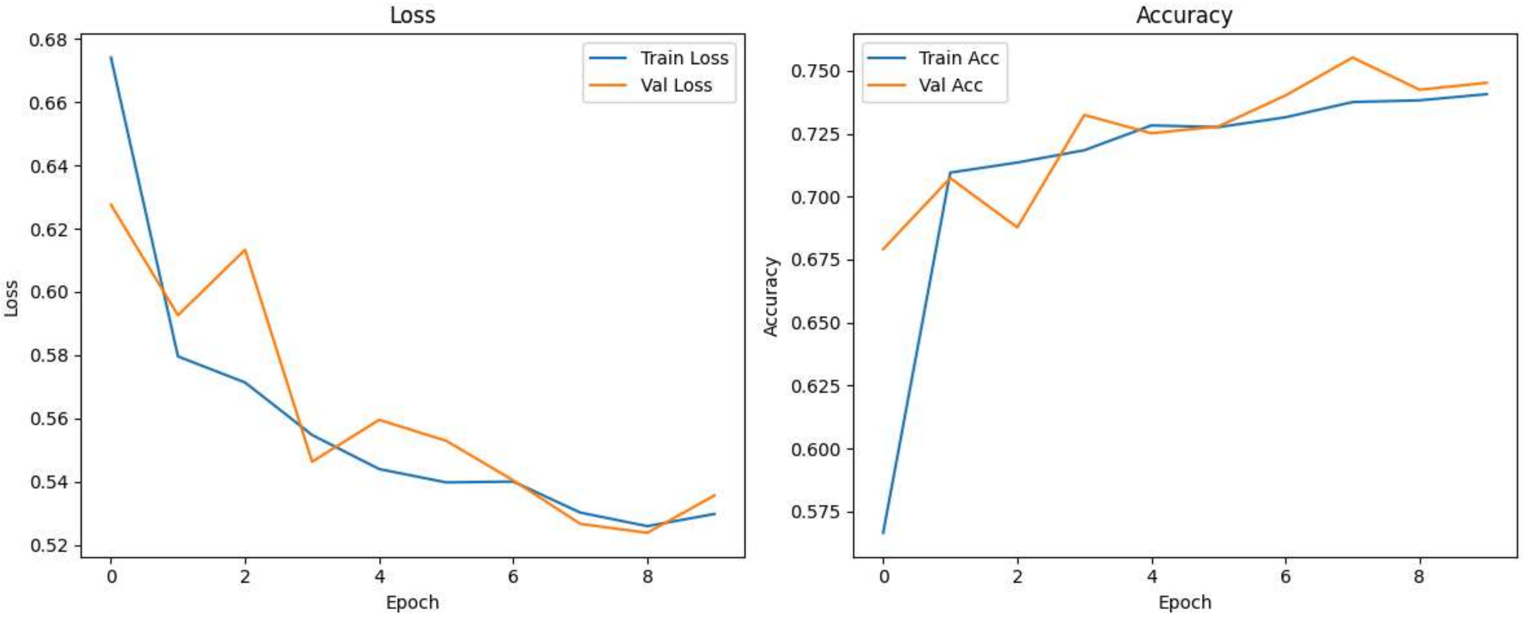


Figure 7

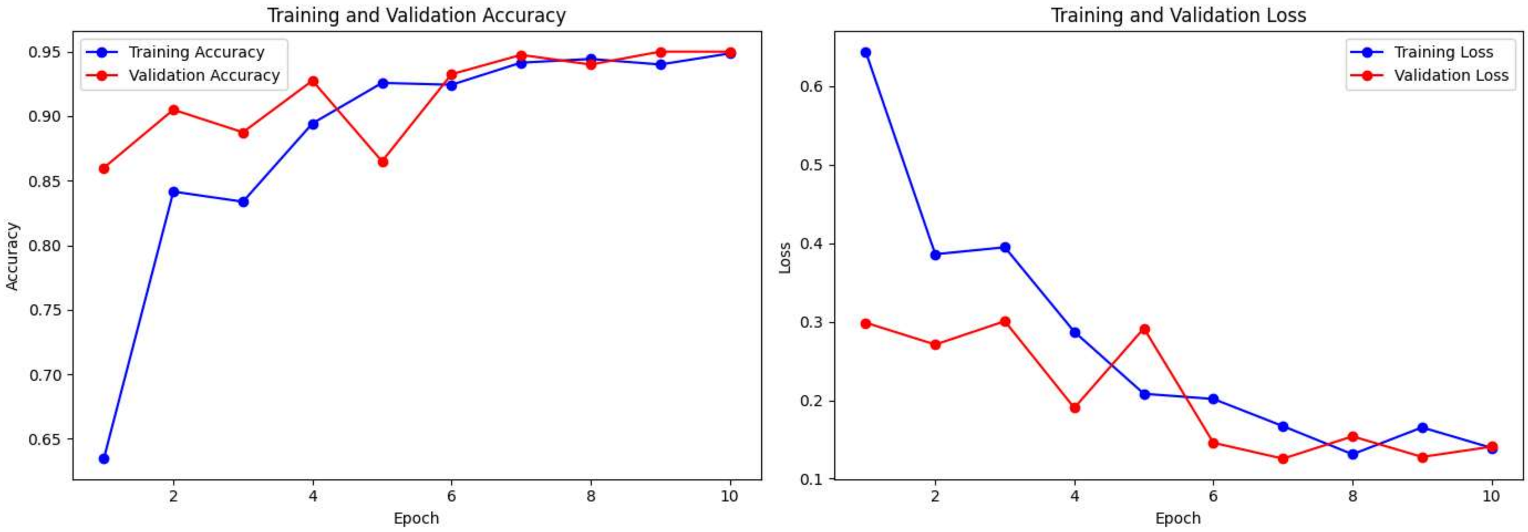


Figure 8

# 5 Conclusion

In conclusion, using CNN or Random Forest-based models, it is evident that they can identify features of galaxy images alone. Further work needs to be done to identify all of these features, and even more to identify more of the classifications that Galaxy Zoo 2 provides. Throughout the testing, I found the most success with the Shape classification, achieving a 95% accuracy. More work needs to be done to build a more complete model, but this has shown what is possible.

# 6 Works Cited

1. Hart, Ross E., et al. “Galaxy Zoo: Comparing the Demographics of Spiral Arm Number and a New Method for Correcting Redshift Bias.” Monthly Notices of the Royal Astronomical Society, vol. 461, no. 4, Oxford University Press (OUP), July 2016, pp. 3663–3682. Crossref, doi:10.1093/mnras/stw1588.
2. Willett, Kyle W., et al. “Galaxy Zoo 2: Detailed Morphological Classifications for 304,122 Galaxies from the Sloan Digital Sky Survey.” arXiv, 2, arXiv, 2013, doi:10.48550/ARXIV.1308.3496.
3. Galaxy Zoo 2 Decision Tree https://data.galaxyzoo.org/gz_trees/gz_trees.html