Deep Learning and Morphology Across Redshifts

Sanchith Shanmuga[1], Soham Patel[1] and Shyamal Mitra[1]

[1]The University of Texas at Austin

ABSTRACT

This paper addresses the question, "How does galaxy morphology differ across red shifts?" Interestingly enough, astronomers can peer through time simply by searching deeper into the universe for galaxies, as the further back one looks in time, the further back they are looking back in time. We utilized this property of physics to analyze galaxies from millions of years in the past to understand how they are structured. The data collection discussed in this paper analyzes galaxies drawn from databases and the statistics collected when running a Convolutional Neural Network (CNN) that is trained on the data set. A discussion of the distribution of morphologies across redshifts is also presented, drawn from the results of the CNN model. Afterwards, an analysis of our CNN model and various distributions are mentioned with our interpretation of the results. Lastly, a reflection about our answer to the research question is put forward with possible future steps to take.

Research Question

Morphology is a characteristic of galaxies that has fascinated astronomers for millennia. The structure and properties of a galaxy are such simple characteristics, yet so complicated to fully understand. Astronomers have peered up at the sky to view these celestial bodies for years, but only in the modern era did astronomers such as Reynolds attempt to categorize galaxies. Reynolds initiated the conversation of galaxy morphology by determining broad categories where galaxies could fall under based on shape, and Hubble took Reynolds' broad classification one step further by developing a galaxy evolutionary sequence: the Hubble Tuning Fork model. The model was not perfect, and today, astronomers understand the slight inaccuracies in Hubble's tuning fork. However, it is a generally accepted base model for categorizing galaxies.

A fascinating aspect about how light travels in space is that by peering at objects millions of light years away, such as distant galaxies, astronomers can view the light reflected by those celestial bodies from millions of years ago. Utilizing this quirk in astronomy, we can analyze how galaxies looked from a large range of time in the past; this is exactly what our research project intends to accomplish. This research paper attempts to analyze galaxy morphology based on the Hubble Tuning Fork model to help understand how galaxies evolved across the eons. Specifically, we are focused on researching how galaxy morphology differs across redshifts. Rather than simply classifying galaxies based on shape, we wish to understand how galaxy morphologies evolved across time, continuing where astronomers like Hubble left off. Realizing how galaxies developed across time can unlock the mystery behind their various evolutionary habits, and with this knowledge, astronomers can better acknowledge how various galaxies in our night sky came to be, including our very own Milky Way galaxy.

Sub-Question 1: Galaxy Morphology and Neural Networks

Given the breadth of astronomy and other research topics that are more abstruse and convoluted, it leaves these large portions of these galaxy image datasets untouched and lost in the newer images that are being captured. With an exponential growth in data and the lack of well experienced astronomers, it has been difficult to get accurate classifications for these datasets that could then potentially be used for research. With the recent spike in popularity and

practical applications, convolutional neural networks have been applied to labeling these datasets based on the Hubble Sequence, some achieving over 95% accuracy on smaller datasets (around 1000 images) (Dai & Tong 2018). Moreover, these networks are computational inexpensive and already trained/implemented (Resnet50, VGG16, etc.).

For our research, we decided to train a Naive Convolutional Neural Network (CNN) with the Galaxy Zoo 2 classifications and the Sloan Digital SkyServer Data Release 7 images. We compared our naive architecture with the Galaxy10 model from the astroNN python package and intend to test other architectures in the future (Leung & Bovy 2018).

Sub-Question 2: Galaxy Morphology and Redshifts

The second sub-question involving our main research question tackles the question about redshifts head on, as we intend to explore how galaxies' (classified into their Hubble buckets based on morphology) morphologies vary across redshifts. As of right now, we plan to tackle the answer to this question based on the data of thousands of galaxies we currently have on file. This data was utilized to train our CNN model. In the future, we plan to employ our CNN to classify validation galaxies and understand how these galaxies' morphologies vary across redshifts.

Data Sources

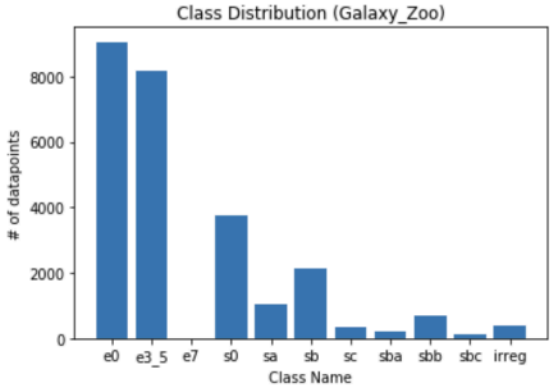


Figure 1. Galaxy Zoo Classification Distribution.

This phase of our research project involves training a CNN on a training set of data and a validation set to ensure accuracy (the data itself being images of galaxies), we required a large amount of galaxy pictures with labels describing the galaxy's classification. Additionally, since we are dealing with classifying galaxies based on morphology through the CNN, the dataset needed to have roughly equal numbers of galaxies per classification category so that the CNN can be trained properly.

We finalized on utilizing the Galaxy Zoo 2 Dataset from the Galaxy Zoo database based on data in the SDSS Data Release 7. The reason we ended up utilizing this dataset is because of the way it is extensive, with data for over 240,000 galaxies, and because of the way it fits our bucketing scheme nicely. This is because of the way the Galaxy Zoo 2 dataset has citizen-gathered values for each step along the decision tree (shown in figure 2), and this makes classifying each galaxy into its respective Hubble bucket relatively straightforward; all we had to do was track a galaxy down the decision tree (and a galaxy moves along the tree based on if it garners over a 50 percent threshold for each step), and at the end of the tree, we place the galaxy into its respective Hubble bucket.

Table 1. Galaxy Zoo compared with expert visual classifications.

GZ2	Galaxy Zoo		HC11		NA10		EFIGI	
	N	(%)	N	(%)	N	(%)	N	(%)
Early-type	79214	86.2	26732	82.1	1995	96.7	214	84.6
Late-type	26314	97.9	79277	88.6	5481	94.9	1675	98.2
Bar	-	-	-	-	651	94.9	238	98.7
Ring	-	-	-	-	438	91.6	110	83.6
Merger	526	63.3	-	-	43	100.	6	100.

In addition to filtering out classifications, we were skeptical of the Galaxy Zoo's classification accuracy given it is done by citizens rather than well trained astronomers. However, after comparing the Galaxy Zoo classifications with expert classifications such as the Nair and Abraham's classifications as well as EFIGI's, we realized the classifications were mostly accurate (Willett et al. 2013). The comparison with expert classifications is demonstrated in table 1. Additionally, the number of irregular galaxies was also small, consistent with general findings about galaxy morphology.

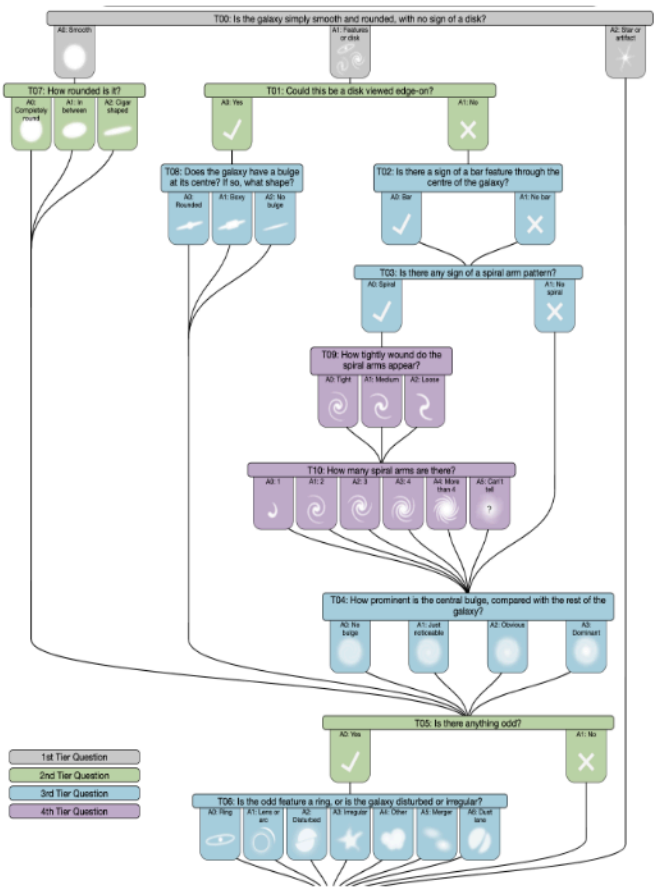


Figure 2. Galaxy Zoo Decision Tree.

Before we were able to train our CNN, we spent a significant amount of time preprocessing and reducing our data. While the Galaxy Zoo 2 contains over 240,000 classified galaxies, a plethora of the galaxies were classified by a significantly small number of people; consequently, it makes it difficult to validate the classifications and could potentially lead to our model establishing unwarranted patterns. From the 240,000 classifications, we only considered those with at least 48 individual people classifying each one and only assigned galaxies to classes if at least 50% of the people agreed in the respective galaxy's classification.

Along with confirming the validity of the Galaxy Zoo's classifications, we constructed a naive bucketing algorithm to align the classifications to the Hubble Sequence. We utilized the Galaxy Zoo Decision Tree shown in figure 2 and compiled the traversals that correspond to each Hubble Sequence class. While the natural distribution seems heavily skewed in figure 1, it is actually justified given that there will be more ellipticals than spiral galaxies.

Furthermore, we were unable to determine an effective way of handling galaxies that are being viewed edge on in determining what spiral class they fall under. As shown in the tree, if a galaxy was viewed edge on, it would bypass any questions concerning spirals (which cannot be determined) and thus not be put in any of the spiral galaxy buckets. This would lead to significantly less spiral galaxies and further extrapolate the skewness of the dataset.

Tools Used

We visualized the Galaxy Zoo's labels and distribution of classes with matplotlib and plotly. Along with these data visualization tools, we used pandas and tensorflow to load in the labels as well as construct our Convolutional Neural Network. For our preliminary analysis of our galaxy data, we decided to use the Galaxy10 CNN as well as construct a naive CNN. With two basic architectures, it allowed us to compare the performance and determine which architecture to define as our base model.

Analysis

Convolutional Neural Network Results

With our data preprocessing and reduction complete, we were able to train a convolutional neural network with heavy constraints on the image resolution, batch size, and number of epochs purely due to computational and time constraints. As shown in figure 3A, we decided to scale the images from the SDSS data release 7 to 64x64 pixels, increased the batch size to 64 in order to reduce the number of steps per epoch, and kept the number of epochs at 5 due to time constraints. We believed these constraints were justified as the astroNN package made similar assumptions when testing their Galaxy10 model (Leung & Bovy 2018). Along with these hyperparameter constraints, we utilized 60% of the 32,000 images as training data and 40% as validation (or testing) data to evaluate our CNN's accuracy. We decided to use a categorical cross entropy loss function given there being 11 categories and an Adam optimization for the gradient descent. We trained 4 different models in order to determine the most optimal hyperparameters (learning rate, batch size, epochs) and architecture for our naive model.

A

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 62, 62, 32)	896
max_pooling2d (MaxPooling2D)	(None, 31, 31, 32)	0
dropout (Dropout)	(None, 31, 31, 32)	0
conv2d_1 (Conv2D)	(None, 29, 29, 64)	18496
max_pooling2d_1 (MaxPooling2	(None, 14, 14, 64)	0
dropout_1 (Dropout)	(None, 14, 14, 64)	0
conv2d_2 (Conv2D)	(None, 12, 12, 128)	73856
dropout_2 (Dropout)	(None, 12, 12, 128)	0
flatten (Flatten)	(None, 18432)	0
dense (Dense)	(None, 128)	2359424
dense_1 (Dense)	(None, 128)	16512
dense_2 (Dense)	(None, 128)	16512
dense_3 (Dense)	(None, 11)	1419

Total params: 2,487,115
Trainable params: 2,487,115
Non-trainable params: 0

B

Layer (type)	Output Shape	Param #
input (InputLayer)	[(None, 69, 69, 3)]	0
conv2d_10 (Conv2D)	(None, 69, 69, 8)	224
activation_20 (Activation)	(None, 69, 69, 8)	0
conv2d_11 (Conv2D)	(None, 69, 69, 16)	1168
activation_21 (Activation)	(None, 69, 69, 16)	0
max_pooling2d_5 (MaxPooling2	(None, 17, 17, 16)	0
flatten_5 (Flatten)	(None, 4624)	0
dropout_10 (Dropout)	(None, 4624)	0
dense_15 (Dense)	(None, 256)	1184000
activation_22 (Activation)	(None, 256)	0
dropout_11 (Dropout)	(None, 256)	0
dense_16 (Dense)	(None, 128)	32896
activation_23 (Activation)	(None, 128)	0
dense_17 (Dense)	(None, 10)	1290
output (Activation)	(None, 10)	0

Total params: 1,219,578
Trainable params: 1,219,578
Non-trainable params: 0

Figure 3. 1st Naive CNN Model Architecture Compared with Galaxy10 Architecture. Figure 3A shows Naive CNN Model Architecture. Figure 3B shows Galaxy10 Model Architecture.

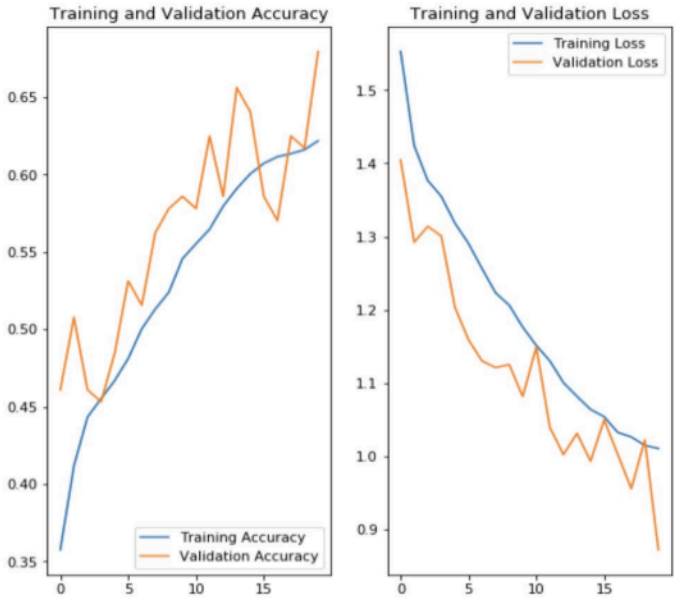
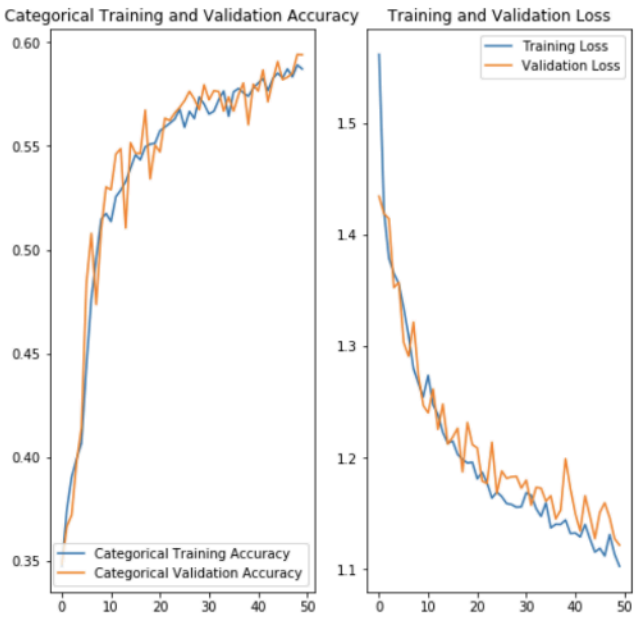


Figure 4. 1st Naive CNN Model Accuracy/Loss Graphs for 20 epochs.

For our first model, we decided to restrict the model's capacity (or trainable parameters) to get some starting accuracies to work off of. Our basic model achieved a training accuracy of 61.1537% and validation accuracy of 67.4532%.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 126, 126, 32)      896
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 63, 63, 32)        0
_________________________________________________________________
dropout (Dropout)            (None, 63, 63, 32)        0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 61, 61, 64)        18496
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 30, 30, 64)        0
_________________________________________________________________
dropout_1 (Dropout)          (None, 30, 30, 64)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 28, 28, 128)       73856
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 14, 14, 128)       0
_________________________________________________________________
dropout_2 (Dropout)          (None, 14, 14, 128)       0
_________________________________________________________________
flatten (Flatten)            (None, 25088)             0
_________________________________________________________________
dense (Dense)                (None, 128)               3211392
_________________________________________________________________
leaky_re_lu (LeakyReLU)      (None, 128)               0
_________________________________________________________________
dropout_3 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_1 (Dense)              (None, 128)               16512
_________________________________________________________________
leaky_re_lu_1 (LeakyReLU)    (None, 128)               0
_________________________________________________________________
dropout_4 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_2 (Dense)              (None, 128)               16512
_________________________________________________________________
leaky_re_lu_2 (LeakyReLU)    (None, 128)               0
_________________________________________________________________
dropout_5 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_3 (Dense)              (None, 11)                1419
=================================================================
Total params: 3,339,083
Trainable params: 3,339,083
Non-trainable params: 0
_________________________________________________________________
None
Found 11921 images belonging to 11 classes.
Found 6430 images belonging to 11 classes.
```

Figure 5. 2nd Naive CNN Model Architecture.




```
Categorical Training Accuracy: 58.72655510902405 Categorical Validation Accuracy: 59.409189224243164
Training Loss: 1.1028275780235008 Validation Loss: 1.1218050211668014
```

Figure 6. 2nd Naive Model Accuracy/Loss Graphs for 50 epochs

For our second model, we increased the input resolution to 128x128 pixels and added an additional pooling layer to reduce the number of trainable parameters to 3.33 million. Additionally, we increased the learning rate to 0.0075. We achieved a training accuracy of 58.7266% and validation accuracy of 59.409%.

Layer (type)	Output Shape	Param #
conv2d_108 (Conv2D)	(None, 126, 126, 32)	896
max_pooling2d_108 (MaxPoolin	(None, 63, 63, 32)	0
dropout_137 (Dropout)	(None, 63, 63, 32)	0
conv2d_109 (Conv2D)	(None, 61, 61, 64)	18496
max_pooling2d_109 (MaxPoolin	(None, 30, 30, 64)	0
dropout_138 (Dropout)	(None, 30, 30, 64)	0
conv2d_110 (Conv2D)	(None, 28, 28, 128)	73856
max_pooling2d_110 (MaxPoolin	(None, 14, 14, 128)	0
flatten_36 (Flatten)	(None, 25088)	0
dense_144 (Dense)	(None, 128)	3211392
leaky_re_lu_108 (LeakyReLU)	(None, 128)	0
dropout_139 (Dropout)	(None, 128)	0
dense_145 (Dense)	(None, 128)	16512
leaky_re_lu_109 (LeakyReLU)	(None, 128)	0
dropout_140 (Dropout)	(None, 128)	0
dense_146 (Dense)	(None, 128)	16512
leaky_re_lu_110 (LeakyReLU)	(None, 128)	0
dense_147 (Dense)	(None, 11)	1419

Total params: 3,339,083
Trainable params: 3,339,083
Non-trainable params: 0

Figure 7. 3rd Naive CNN Model Architecture.

Categorical Training Accuracy: 62.494730949401855
Categorical Validation Accuracy: 60.96450090408325
Training Loss: 1.0292753430797066 Validation Loss: 1.098958559036255

Figure 8. 3rd Naive CNN Model Accuracy/Loss Graphs for 50 epochs.

For our final model, we adjusted the learning rate to 0.00595 to improve the accuracy from the third model. We achieved a training accuracy of 62.4947% and a validation accuracy of 60.9645%.

Analysis of Galaxies Across Redshifts

Our sub-question for this section is specifically, "How are galaxy morphologies distributed across various redshifts?" Currently, we were able to query and create a large dataset of around 32,000 galaxies and various characteristics about them, such as their location (ra and dec) as well as their redshifts and size in the sky (this metric is useful when obtaining images of the galaxies from SkyServer to train our CNN). We will attempt to answer this question based on this 32,000-galaxy dataset, which is the same dataset we utilized to train our CNN. In the future, we hope to gain further insight to this redshift question by retrieving random images of galaxies for our CNN to classify. However, we currently have a strong conclusion to draw based on the large dataset we analyzed.

We began with a set of 245,000 galaxies obtained from the Galaxy Zoo 2 dataset, and our data reduced this to only include galaxies that were classified by more than 48 people.

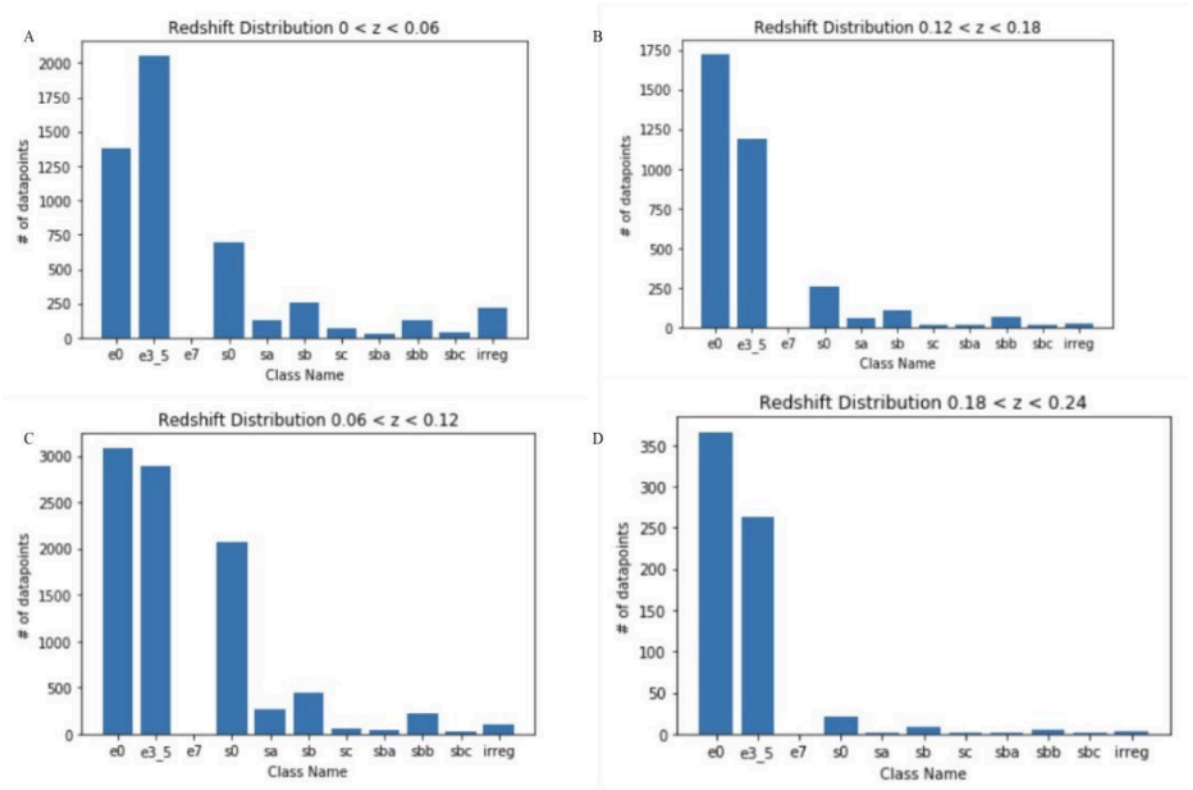


Figure 9. Graphical Representation of Galaxy Morphology Across Redshifts. Figure 9A shows galaxy morphology distribution between redshifts of $0 < z < 0.06$. Figure 9C shows galaxy morphology distribution between redshifts of $0.06 < z < 0.12$. Figure 9B shows galaxy morphology distribution between redshifts of $0.12 < z < 0.18$. Figure 9D shows galaxy morphology distribution between redshifts of $0.18 < z < 0.24$.

Figure 9 represents how galaxies' morphologies differ across four different redshift ranges in a graphical format. The galaxies are drawn from our data-reduced dataset of 32,000 galaxies and are split into four approximately equal, but arbitrary redshift ranges. The upper and lower bounds of the redshifts are determined simply by the galaxies in our dataset and their redshifts.

The results from this distribution across redshifts reveals a curious aspect: the distribution of galaxy morphology seems to be consistent more or less across the various redshift ranges. In each of these redshift ranges, spiral galaxies of class E0 and E5 tend to dominate, spiral galaxies (especially non-barred galaxies) are the next most prominent galaxy types, and there seems to be around the same number of SBa, SBc, and irregular type galaxies. The SBb class of barred galaxies unusually is found in greater prominence across all the redshifts compared to the other barred galaxy types.

Interpretation of Results

Interpretation of CNN

Even though the CNN accuracies imply that not only did our model perform extremely well, it also outperformed the Galaxy10 model, we suspect these are due to the input image constraints as well as the slight differences in the model architectures. As shown in figure 3A our model had 2,487,115 parameters while the Galaxy10 model (seen in figure 3B) had 1,219,578 parameters; consequently, the improved performance could be directly correlated with simply having a more convoluted, ”deeper” model as that is the only difference between the two model architectures. Going forward, we would like to implement other models such as the R-CNN described in figure 10 and other pre-built models such as the VGG16 and Resnet50 to determine what attributes of the model directly correlate to improved accuracy. Although there are a plethora of models to test, most of the models mentioned above require significant computational power as well as time to even train smaller datasets; consequently, we decided to only compare our naive model and the Galaxy10 given the computational and time constraints we imposed.

Table 2. Comparative analysis of naive models trained for 50 epochs.

Model	Training Accuracy	Validation Accuracy	Training Loss	Validation Loss	Trainable Parameters	Learning Rate
1	61.1537%	67.4532%	1.0149	0.8468	2,487,115	0.01
2	58.7266%	58.4092%	1.1028	1.1218	3,339,083	0.0075
3	62.4947%	60.9645%	1.0292	1.0990	3,339,083	0.00595

One of these constraints was our low image resolution. With a 64x64 pixel input image, it is definitely possible that the model identified certain patterns/irregularities that wouldn’t be visible in higher resolution images; moreover, given the nature of galaxy images, there would be significantly more pixels that are black and thus useless. While humans know to ignore these pixels, our model weighs all these pixels the same, leading to the model incorporating the black pixels into classifying new galaxies rather than just considering the patterns.

While the input image resolution does heavily impact our model’s performance, it also implies that the model's accuracy could increase further with more tuning and convoluted models. With higher image resolutions, the model will be able to generate convolutions that extract more details and thus increase the number of parameters

utilized to classify the image. To test this, we decided to train 3 separate models with various learning rates, architectures, and trainable parameters to determine the optimal architecture for our naive model after 50 epochs. The results for each model is displayed in table 2 and displays the impact of the learning rate and trainable parameters on the training and validation accuracy.

While Model 2 did utilize a higher input resolution, we added an additional pooling layer to reduce the number of trainable parameters to 3.33 million. In addition to lowering the number of trainable parameters, we also increased the learning rate from 0.001 to 0.0075 to introduce the implicit regularization a higher learning rate brings. By addressing both these problems, we were able to achieve a training accuracy of 58.727% and validation accuracy of 58.4092%, significantly reducing the difference in accuracies and the model's overfitting tendencies. We concluded that the low accuracies were likely due to the learning rate being too high, preventing the model from finding the global minimum in the loss functions, and these claims can be corroborated with figure 8, as the losses are clearly fluctuating and still seem to be decreasing. In order to address the low accuracies of Model 2, we adjusted the learning rates in Model 3 for our final naive CNN model.

In Model 3, we decided to utilize the same architecture as Model 2 as there were no signs of overfitting in Model 2; however, we decreased the learning rate from 0.0075 to 0.00595 in order to enable our model to minimize the loss. We were able to achieve a training accuracy of 62.4947% and a validation accuracy of 60.9645%. Though the accuracies diverge slightly, we believe the improvement in accuracies outweigh the divergence and determined this is the most optimal naive model and hyperparameters for it. As the loss functions still seem to be decreasing, we predict that with more epochs the model will be able to achieve significantly higher accuracies and lower losses.

Interpretation of Morphology Across Redshifts

One key process to note when analyzing the results of our redshift distributions is the significant data reduction we did with our datasets. Data reduction was necessary for two reasons: accuracy and efficiency. Because Galaxy Zoo is a crowdsourced dataset and not one crafted by experts, some of the classifications may be incorrect. To minimize this from altering our results significantly, we concluded to only consider galaxies that have been classified by more than 48 people. This promises some more stability in our classifications. Additionally, 48 was a reasonable number for our purposes because it resulted in a large enough dataset (32,642 galaxies to be exact) that also did not result in extremely long querying times when we had to query for redshift values.

After performing the data reduction, we queried the redshift value for each of the 32,642 galaxies since the dataset did not come with that value beforehand. This was a timely operation, as it took a few hours to pull all these values from the SDSS Data Release 7. In addition, we also queried the images for each of the galaxies from the SkyServer and since each galaxy's image had to be queried individually, for all the galaxies in our dataset, this also took approximately 3 hours to complete. The images were utilized to train the CNN by classifying the 32,642 galaxies into the various Hubble buckets employing our bucketing scheme.

After the CNN was trained with these images and all the galaxies were placed into their respective Hubble buckets, we graphed how the morphologies of the galaxies are distributed in four different bands of redshifts, which can be seen in figure 9.

Scientific Value of Research

Previously, we stated that our conclusion based on the results was that there is no significant difference in the morphological distribution of galaxies across redshifts; factoring a change in the scale, the distribution of galaxies remained the same. What this could mean, for the redshifts that we have data for, is that from an evolutionary perspective, galaxies have been shaped the same across time. By this, we are specifically inferring that galaxy morphological type is independent from time; throughout time, for example, elliptical galaxies have dominated over other galaxy types such as the various spiral types and irregular types, and this continues to be the case today. This can help reveal

some errors in Hubble's tuning fork evolutionary model as well, as Hubble proposed that galaxies evolved from elliptical types into barred or non-barred spiral types. If Hubble was correct, then transitioning from high to low redshifts, we would see less elliptical and more spiral galaxies. However, based on the data in figure 9, the number of elliptical galaxies stays dominant over the number of spiral galaxies, and the distribution of morphologies stays consistent over redshifts. This gives evidence against Hubble's evolutionary tuning fork and can corroborate modern astronomers who have reached a consensus that Hubble's evolutionary model is incorrect.

However, one important difference to note is that the number of galaxies in each morphological type generally decreases as we examine larger redshifts. This could indicate a timetable when it comes to galaxy formation; in larger redshifts (such as $0.18 < z < 0.24$), galaxy formation was more dormant as opposed to closer redshifts (such as $0 < z < 0.06$ and $0.06 < z < 0.12$). This timeframe can be useful when understanding how our universe evolved as a whole and how it looked eons ago. For example, the redshifts of $0.18 < z < 0.24$ could portray an emptier universe of fewer galaxies, but eventually led to a universe with more abundant galaxy formation. However, we cannot make this claim with confidence, because the less number of galaxies at higher redshifts can simply be a property of the Malmquist bias. This bias occurs when it is difficult to detect high redshift galaxies (even if their absolute luminosity is great) because they are so faint due to their great distance from Earth. Furthermore, the detection of these high redshift galaxies are hindered by the large quantity of galaxies at smaller redshifts. Therefore, the small number of galaxies in our dataset at higher redshifts could simply be the result of the Malmquist bias as opposed to an actual indicator of galaxy formation.

Similar Research

Neural Networks and Morphology

The problem of galaxy classification is a textbook, multiclass image classifier problem. The versatility of convolutional neural networks has enabled astronomers to utilize computational astrophysics to automate galaxy classification with various model architectures: A Masked Region Based Convolutional Neural Network and a Variational AutoEncoder, both using supervised and unsupervised learning respectively to group galaxies together.

The Masked Region Based Convolutional Neural Network (R-CNN) utilized supervised learning, or where the data (our galaxy images) already had predefined classes that each one needed to be put into. The model's architecture can be broken down into two sections: the convolution layers and the classifier (multi-layer perceptron). The predefined classes came from the Hubble Sequence and the model is Region based because it trims the images around the Regions of Interest. Many of these galaxy images contain a significant number of black pixels, which are unnecessary and can be ignored when trying to classify galaxies. Since this model's efficiency is dependent on the input image resolution, cutting down as many pixels as possible will draw "focus" towards the galaxy rather than extraneous features in the background (increasing accuracy and validity of the model). The image will then pass layers of convolution and pooling before each pixel serves as an input parameter for the classifier. After the pixel goes through several layers of nodes, it will output probabilities which represent which class it believes the galaxy is. H. Farias, the author of the paper, used the Galaxy Zoo dataset in order to train this model, achieving 92% accuracy on the testing set (Farias et al. 2020). This model's architecture will serve as an incredible starting point to tune for our personal model for galaxy classification.

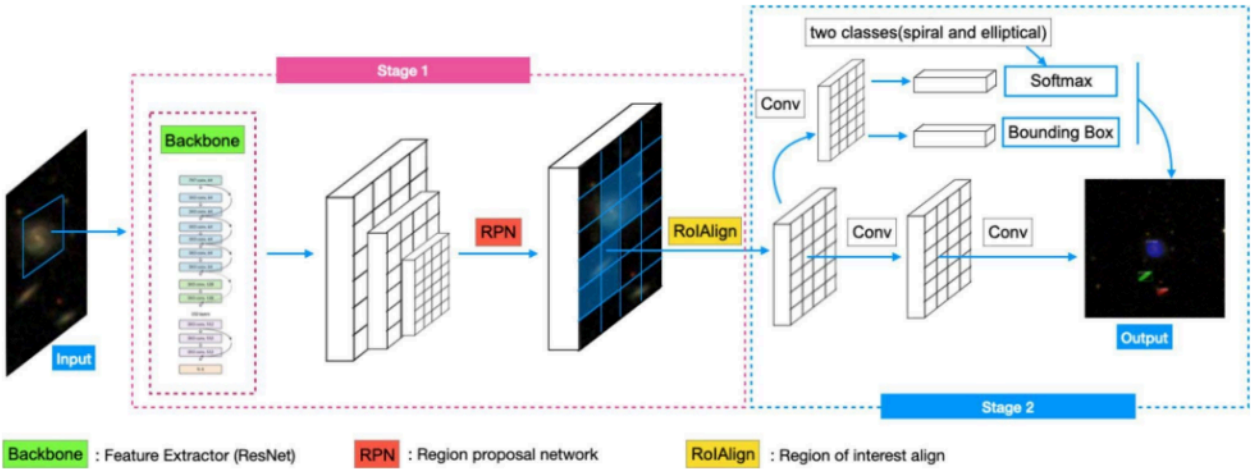


Figure 10. Masked R-CNN Architecture from Farias et al. (2020).

Along with the R-CNN, there was another model which utilized unsupervised learning, where the classes are not predefined, to cluster galaxies as well as generate synthetic images of galaxies for larger datasets. The AstroVaDer employed a variational autoencoder (VAE) to cluster galaxies into classes based on their features alone, not any external classification model. Ashley Spindler claimed they chose to use unsupervised learning given the nature of the galaxy data as well as the severe limitations of supervised learning (Spindler et al. 2020) Galaxy images are largely unlabeled, making it extremely difficult for researchers/astronomers to acquire largely labeled datasets and almost invalidates Convolutional Neural Networks given the relatively smaller datasets. AutoEncoders learn how to encode data into a lower dimension representation (latent space) and decode it back to its original dimensional space (Spindler et al. 2020). Not only does this not require any labels to create the clusters of galaxies, but it also can generate synthetic, realistic galaxy images for other datasets/models. While VAEs are extremely powerful in transforming data from embeddings to its original form, Generative Adversarial Networks are said to perform significantly better. However, in spite of unsupervised learning not requiring labeled data, it also prevents us from adhering to the Hubble Sequence. These unsupervised models could potentially lead us to a completely new classification model, but also prevents us from exploiting the deductive capabilities of the Hubble Sequence.

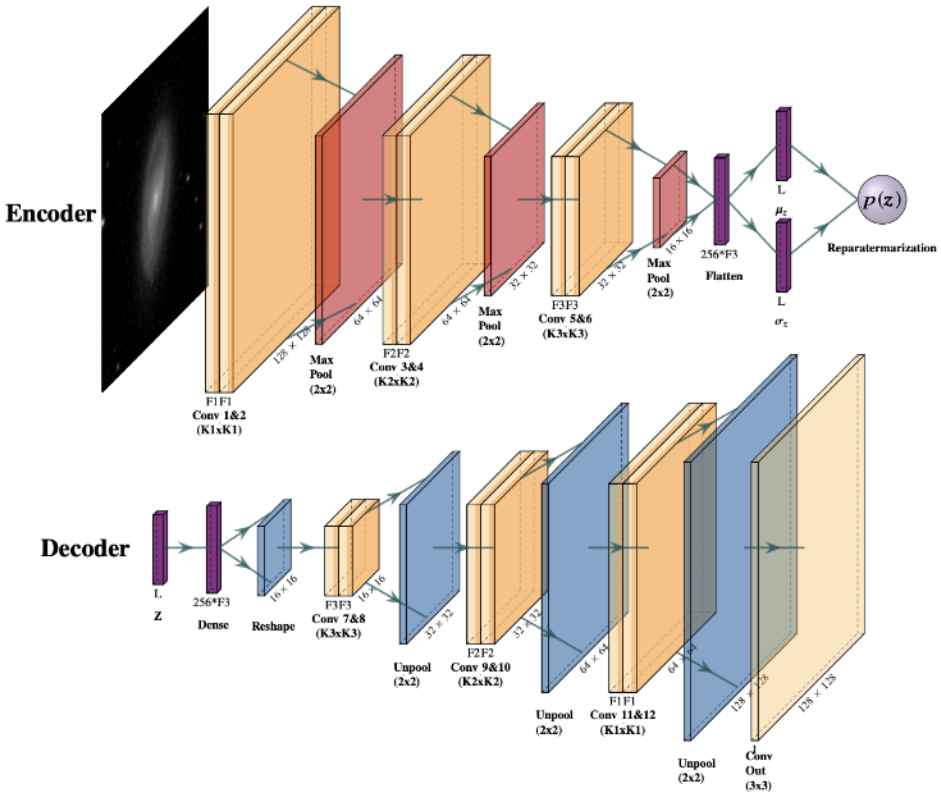


Figure 11. VAE Architecture from Spindler et al. (2020).

These models and papers justify the applications of neural networks and machine learning in astronomy, specifically in galaxy morphology. However, one thing these models lacked was following the Hubble Sequence classification model for their output. Not only do these papers emphasize the breadth of models that can be trained and applied, it provides us with numerous opportunities to test our own naive model against to determine the most effective model through comparative analysis.

Galaxy Morphology and Redshift

The most significant study we discovered regarding galaxy morphology across redshifts from outside sources came from a study done by Conselice where they had an interesting, and somewhat differing conclusion: they discovered that, “At higher redshifts (early times) the fraction of bright galaxies that are peculiar in structure and morphology increases gradually at the expense of both spirals and ellipticals” (Conselice 2004).

Essentially, the researchers are making the claim that spiral and elliptical galaxies are found less frequently at higher redshifts as opposed to “peculiar” structured galaxies, which we are assuming to be irregular galaxy types. Our distribution of morphologies across redshifts, as seen in this figure 9, did corroborate the external researchers’ claim that elliptical and spiral galaxies decreased in number at higher redshifts as we found a smaller number of those galaxies then, but it does not seem to corroborate the claim that irregular galaxies increased at higher redshifts. The number of irregular galaxies, according to our data, actually decreased as higher redshift values were surveyed. However, this difference in the conclusion could also be a result of the amount of galaxies we surveyed; if we collected data for hundreds of thousands more galaxies at various redshifts, we might see our conclusion match the external researchers’ conclusion better.

Reflection and Conclusion

Throughout this project, we were able to thoroughly preprocess and reduce our dataset, applying our statistics knowledge to ensure that the dataset used to train our model was accurate and had minimal inherent bias through certain assumptions made. We compared the Galaxy Zoo 2 dataset with expert classifications, utilized the debiased fraction of votes per question, and only considered galaxies with statistically significant classifications in order to ensure our model learned to identify specific features rather than random noise unique to the dataset. While we are confident that our data preprocessing was thorough, it prevented us from performing other analyses with the model as well as determining a better bucketing method.

Though our results were satisfactory for two models, we realized we were unable to do comparative testing with other models such as the Masked R-CNN or the AutoEncoder. These comparisons could have helped us determine a more effective and optimal model for the dataset, potentially resulting in better performance and a more convincing result. Additionally, the inability to address edges on galaxies from the Galaxy Zoo lead to a more skewed class distribution.

We trained our CNN with a dataset of around 30,000 images, and we are satisfied with the results our dataset gave for our CNN accuracy and resulting distributions. Providing our CNN with thousands of galaxy images and being able to draw conclusions about morphology across redshifts is only feasible with a large dataset, which we believe we did. We are also complacent with how we were able to problem solve when querying redshift values and galaxy images from the SDSS and SkyServer, as these operations resulted in a lot of errors and time. However, we were able to successfully obtain this data to train our CNN and understand the distribution of morphologies across redshifts.

However, what we could have done further to better answer our research question would be to train our CNN with an even larger dataset, with the number of galaxy images on the order of around 200,000. This may help us develop a more accurate CNN model to test validation data with. However, we were constrained by time and computing power to accomplish this feat, and in the future, if these issues are solved, we may be able to further improve our results.

Additional Work

For future work, we would like to increase the number of galaxies used to train our model by reducing the constraints we had performing our data reduction on the Galaxy Zoo 2 dataset. For example, rather than only considering galaxies with more than 48 classifications, we could consider galaxies with 40 classifications, increasing the number of datapoints from 32,000 to 100,000 and thus increasing our dataset size.

In addition to expanding our dataset size, we would like to explore other models such as a Bayesian CNN and Unsupervised learning for labelling as these models are better fits given the nature of our dataset. The Bayesian CNN allows the model's output to incorporate uncertainty and probabilistic morphology, which better fits the Galaxy Zoo 2 dataset as all the classifications given are as percentages rather than absolutes. Additionally, we could significantly improve the labeling we used for the Galaxy Zoo 2 Dataset to incorporate unsupervised learning, generating more accurate labels and allowing us to utilize a large portion of the dataset.

Acknowledgements

We would like to thank Dr. Shyamal Mitra for his support, guidance, and assistance in all of our work during the last year in the Freshman Research Initiative program. We would also like to extend our thanks to our teaching assistants Zhimin and Harsh for instructing us in the fundamentals of Python and data science. Additionally, we would like to thank our peers in the Geometry of Space stream for all of their help and support in assignments and projects over the

last year. Lastly, we would like to thank Dr. Mitra, Dr. Karl Gebhardt, the Freshman Research Initiative program, and The University of Texas at Austin for their work in establishing and maintaining this research stream as an avenue for undergraduates to learn about the world of scientific research.

References

Conselice, C. J. 2004, Penetrating Bars through Masks of Cosmic Dust, 489–510, doi: 10.1007/978-1-4020-2862-544

Dai, J.-M., & Tong, J. 2018, Galaxy Morphology Classification with Deep Convolutional Neural Networks. https://arxiv.org/abs/1807.10406

Farias, H., Ortiz, D., Damke, G., Jaque Arancibia, M., & Solar, M. 2020, Astronomy and Computing, 33, 100420, doi: 10.1016/j.ascom.2020.100420

Leung, H. W., & Bovy, J. 2018, Monthly Notices of the Royal Astronomical Society, doi: 10.1093/mnras/sty3217

Spindler, A., Geach, J. E., & Smith, M. J. 2020, arXiv e-prints, arXiv:2009.08470. https://arxiv.org/abs/2009.08470

Willett, K. W., Lintott, C. J., Bamford, S. P., et al. 2013, Monthly Notices of the Royal Astronomical Society, 435, 2835, doi: 10.1093/mnras/stt1458