Draft version December 12, 2022
Typeset using LaTeX default style in AASTeX631

Astron 128 - Lab 3: Morphological Classification of Galaxy Zoo Dataset

Shrihan Agarwal[1]

[1] *University of California, Berkeley, Department of Astronomy, Berkeley, CA 94720*

Abstract

The future of our understanding of astronomy lies in large-scale spectroscopic surveys. With immensely large astronomical datasets, comes the need for algorithms that can sift through and classify galaxies, and identify interesting features for further investigation. The Galaxy Zoo 2 dataset (Lintott et al. 2008; Willett et al. 2013), is one such dataset with visually inspected morphological galaxy classifications providing insight into formation and subsequent evolution of galaxies. We utilize a ResNet50 convolutional neural network schema to classify over 60,000 Sloan Digital Sky Survey galaxy images. Trained on an initialization of ImageNet-based weights, the ResNet50 model achieves a good validation accuracy, with a root mean squared error of 0.107 overall for all model classifications. We describe further work that may be done to optimize this result with a greater time, computational capacity, and variation of hyperparameters.

1. INTRODUCTION

The Galaxy Zoo project was launched in 2007 as an opportunity to involve citizen science to provide morphological classifications for nearlry one million galaxies in the Sloan Digital Sky Survey (Lintott et al. 2008). The success of the initial project, which classified galaxies into elliptical (early-type), spiral (late-type) and merging galaxies, prompted a larger, more in-depth classification, as described in Willett et al. (2013). In brief, the classification scheme now includes support for artifacts, and increased information regarding the number of spiral arms, barred and unbarred spirals, and the shape and inclination of the elliptical galaxy.

Though machine learning has been applied to astronomy for quite some time, the recent growth in the computational ability of computers, particular focus on the development of neural networks, and the rise of large-scale sky survey datasets have caused an expansion of its use (Sreejith et al. 2018; Metcalf et al. 2019), both in galaxy classification and other areas like the strong gravitational lens search challenge.

The understanding of galaxy morphology and structure is critical to our understanding of the formation and evolution of galaxies with time. By analyzing galaxies upto z = 3, we can judge the galaxy evolution history of the universe in terms of their concentration, merger rates, and more. For example, we can find that galaxies are 2-5 times smaller at higher redshift, for a given stellar mass (Conselice 2014).

We briefly describe our dataset in Section 2. We begin discussing preprocessing, inspecting the classification labels and analyzing correlations in 3. Following which, we descirbe the results of our ImageNet-based classification model 4. Lastly, we present our conclusions and possibilities for further improvement in Section 5.

2. DATASET

The Galaxy Zoo 2 dataset consists of 61, 578 424 x 424 pixel images on the sky, with an pixel scale of 0.396 arcsec/pixel, resulting in an image angular diameter of approximately 167 arcseconds. The sample of galaxies is selected from the brightest 25% of galaxies in the SDSS Northern Galactic Cap region, and the SDSS DR7 catalogue specifically. Multiple cuts were performed to curate this dataset - namely, providing us with the nearest, largest,

Corresponding author: Shrihan Agarwal
shrihan@berkeley.edu

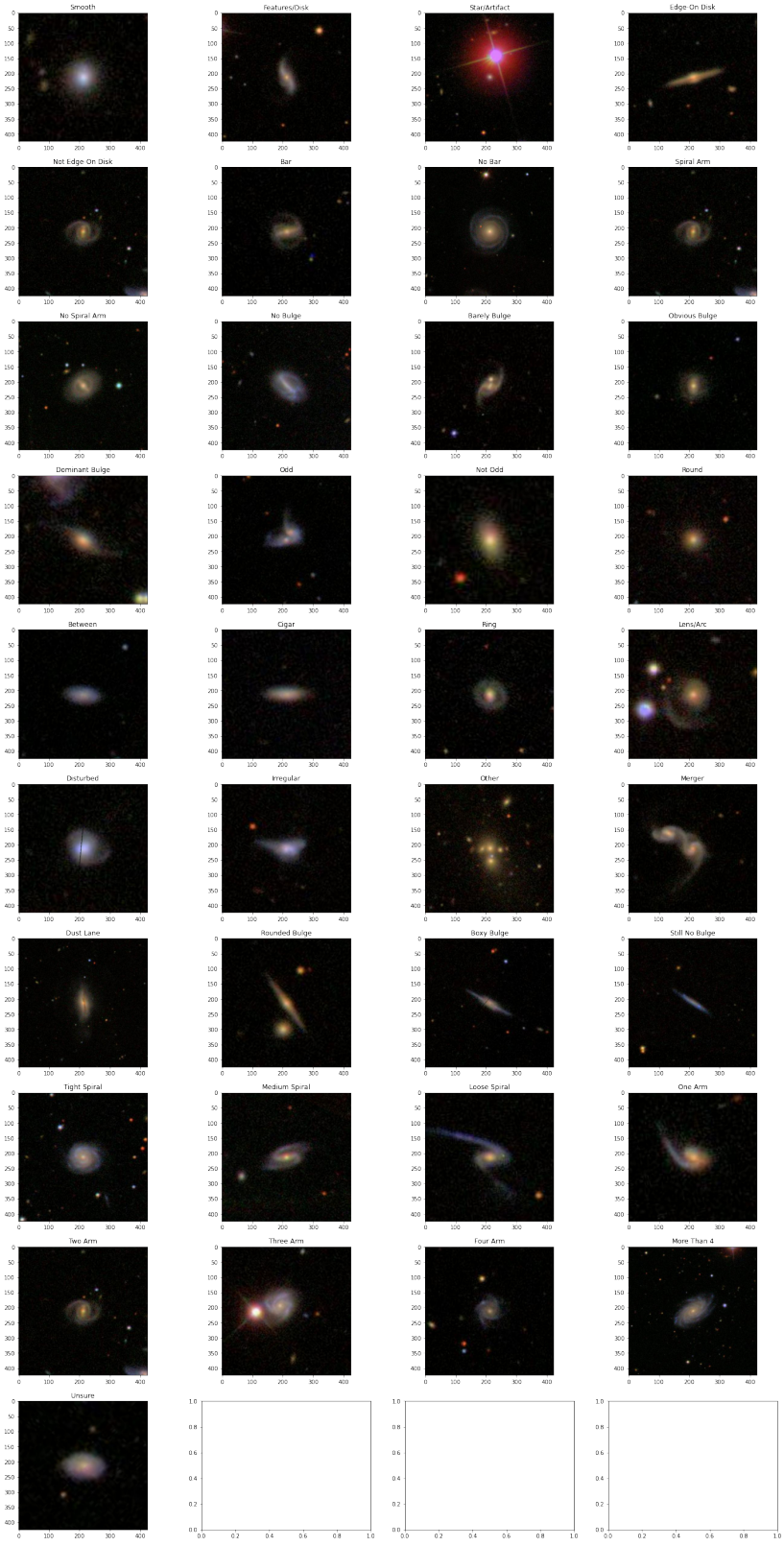

Figure 1: Examples of each of the 37 different classification responses allowed under the GZ2 classification scheme.

and brightest systems for accurate galaxy classification. Spectroscopic redshifts of galaxies greater than 0.25 were removed, and objects flagged as saturated, bright or blended by the pipeline were excluded. The dataset also includes a stripe of galaxies from the SDSS Southern Galactic Cap, along the celestial equator region, which required additional processing as described in Willett et al. (2013).

The GZ2 dataset has 11 decision steps for the user to classify the galaxy, with a total of 37 possible responses. See 1. The human classifier can only select one response. The images shown to the users were randomly selected from the curated database, and towards the later end of the project, images with low numbers of classifications were shown more often, to allow a uniform set of responses. Over 99.9% of the sample had over 28 classifications per image.

In the GZ2 survey, there is a "classification bias", causing a change in observed morphology fractions as a function of redshift, independent of any true galaxy properties. In Willett et al. (2013), an assumption is made that the survey is shallow enough ($z < 0.25$) to justify an assumption of no evolution to the degree seen in the classification scheme. They instead attribute this classification bias to a decrease in the quality of the imaging of smaller, more distant galaxies, where spiral features often blend to appear elliptical. As a result, galaxies with spiral features such as disk galaxies and bar structure decreases with increasing redshift. Merger galaxies become more common as well, since the angular separation between galaxies decreases as a function of redshift, making it appear as though far apart galaxies are actually part of the merger population. These are systematically debiased, assuming that for galaxies of the same brightness and size, a different sample of galaxies with similar brightness and size will share the same average mix of morphologies.

3. DATASET CORRELATIONS AND PREPROCESSING

3.1. *Correlations in Galaxy Classification Responses*

In the visually inspected GZ2 dataset, classification responses are not equally distributed in the training set. Furthermore, some classifications are exclusive of others - for example, a bar and no-bar classification are exclusive, and likely have a strong negative correlation. We take the entire GZ2 image labels, and derive correlations for various parameters in Figure 2.

We discover interesting correlations. A few of note include weak correlations between stars and smooth galaxies, which may appear similar if there is an abscence of diffraction spikes or to amateur astronomers. There is a strong positive correlation between edge on disks and rounded bulges, not as a result of any physical property, but rather the nature of the classification scheme, which asks if a disk has a bulge feature following a classification as an edge on galaxy. We find a correlation of 0.5 between cigar-shaped and edge-on disk galaxies, indicating a degeneracy in the classification scheme, where it is difficult to differentiate between the two different types of classifications.

3.2. *Class-Specific Correlations in GZ2*

The correlations become increasingly clear when considered in the context of specific decision stages in the Galaxy Zoo process. For example, the Class 1 classification, determining if an image is classified as a star, smooth galaxy, or disk galaxy, the correlation of 0.28 between stars and smooth galaxies is noticeable, indicating an uncertainty in their classification. There are also weak correlations between "tight spiral" and "medium spiral", or "medium spiral" and "loose spiral", indicating the subjective nature of the classifications overall. This is also seen between 3 arm and 4 arm spirals, and weakly between bulge size classifications.

The uncertainty of certain classifications is clear from these correlations. A correlation of 0.57 between rounded and boxy bulges indicates the high uncertainty in their classification by visual inspection. Similarly, we find uncertain correlations between disturbed and irregular galaxies, as well as between "Merger" and "Other Odd Feature" classifications. See Figure 5.

3.3. *Generating Image Cutouts*

The GZ2 dataset is large and managing the entire dataset of 65000 4MB files is computationally challenging to feed into a convolutional neural network. We generate image cutouts using the Pillow Image handling library in Python, and downscale the image resolution. The image is first cropped to a central box of 120 x 120 pixels, and

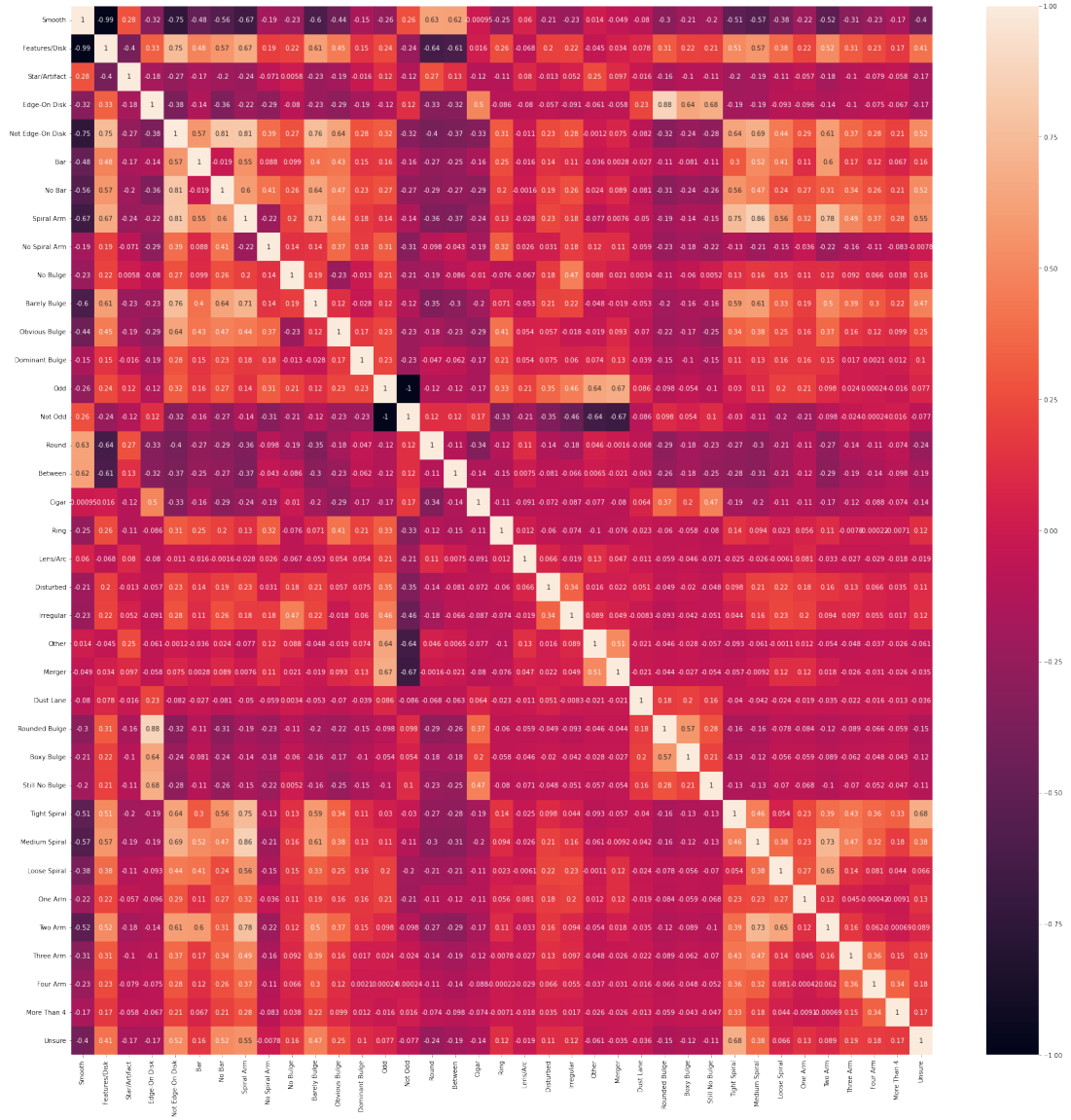

Figure 2: Correlations Among Galaxy Zoo Classification Labels. We find strong, expected negative correlations, such as that between "Odd" and "Not Odd" classifications, or Smooth and Feature classifications. Interestingly, there are also weak correlations between stars and smooth galaxies, and improved detections (higher correlation) of bulges in non-edge on disk galaxies. This demonstrates the nature of the classification, and points to classification biases as a result of the nature of the classification process.

then downsampled to a 70 x 70 pixel grid. The pixels themselves are dropped from a 64-bit floating point precision to a 32-point precision, overall dropping the image size by approximately 80 times. A sample of the cropped images is provided in Figure **??**.

3.4. *Image Preprocessing and Data Augmentation*

The images to be fed into the CNN were optimized for machine learning. The cropped images were rotated by a random angle between 0 and 360, before being normalized on a per image basis and added to the batch for training. This form of data augmentation and image preprocessing prevents overfitting by introducing a random rotation to all of the training data, ensuring that the algorithm cannot "learn" the training images themselves. For our galaxy training dataset, the images are rotationally invariant, as the images can be taken with any orientation in space.Additionally,

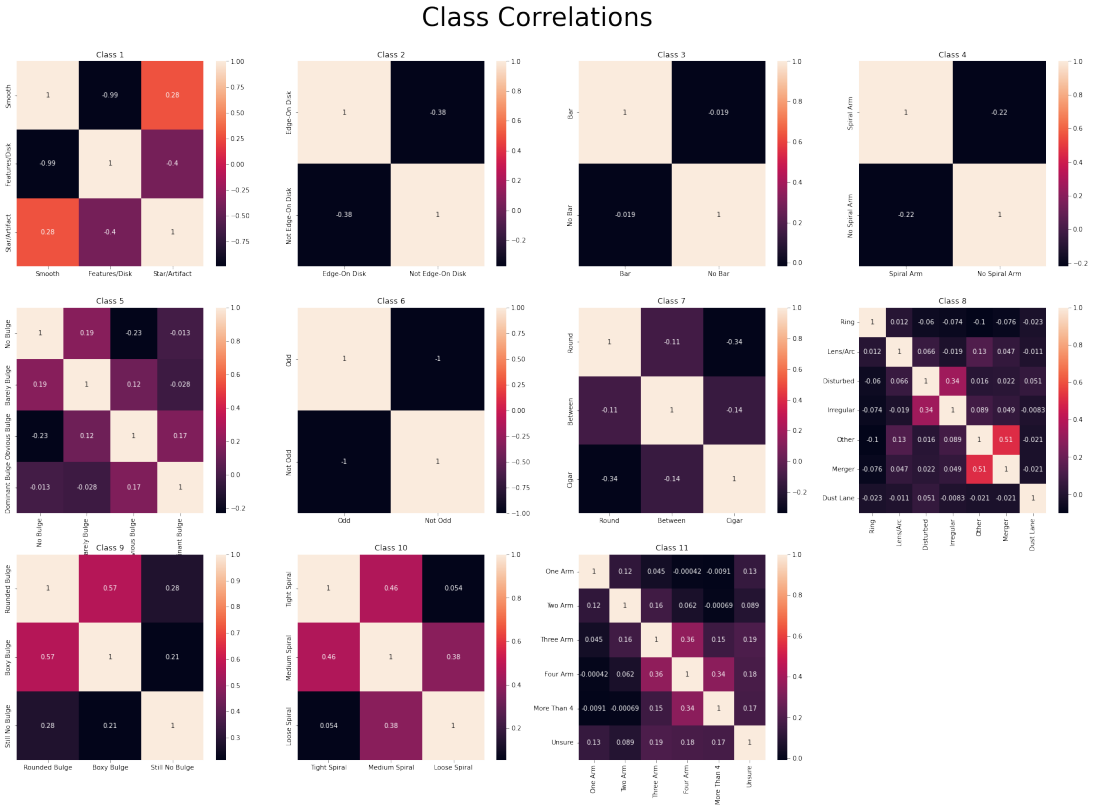

Figure 3: Class-Specific correlations among galaxy zoo labels. These highlight the more uncertain decisions made in the visual inspection dataset, for example between boxy and rounded bulges, merger and "other odd feature" classifications, and disturbed and irregular galaxies. We once again notice the weak correlation between stars and smooth galaxies.

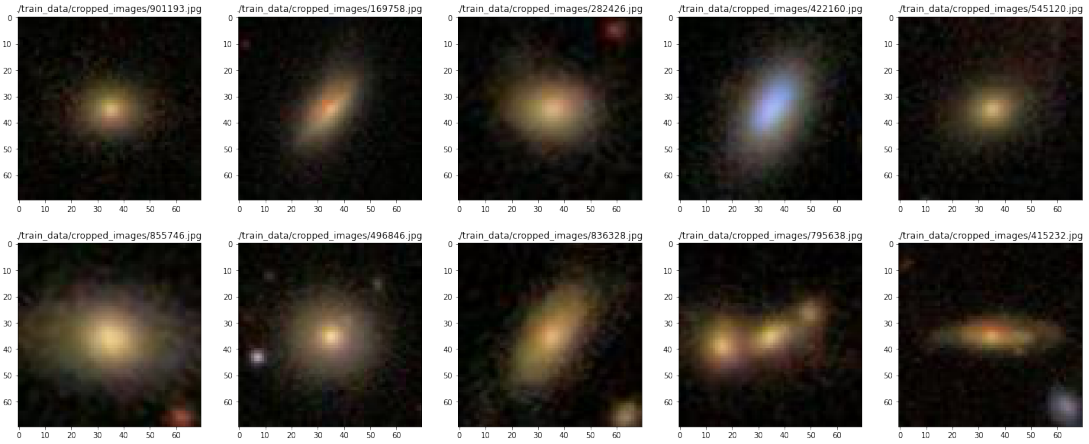

Figure 4: Examples of randomly selected cropped images. The large majority of images retain information necessary for galaxy classification, while the dataset size is reduced by 80 times.

providing the renormalized images to the convolutional network helps it perform better and ensures the trained weights are not excessively large.

3.5. *Rescaling Weights*

The output of the neural network must ensure scaled weights according to the class probabilities as determined by the decision tree approach of GZ2. Specifically, the outputs of the neural network need to be renormalized as the

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 resnet50 (Functional)       (None, 3, 3, 2048)        23587712

 flatten (Flatten)           (None, 18432)             0

 dense (Dense)               (None, 37)                682021

 activation (Activation)     (None, 37)                0

=================================================================
Total params: 24,269,733
Trainable params: 24,216,613
Non-trainable params: 53,120
_________________________________________________________________
```

Figure 5: ResNet layers wrapped to output into a dense layer with 37 classifications corresponding to the 37 possible responses in the GZ2 decision tree. The number of trainable parameters in the model is 24, 216, 613.

decision probabilities are not independent of each other: for a disk galaxy, the edge-on and not edge-on configurations have a dependent probability, for example. We rescale the weights accordingly, renormalizing the probabilities for each set of decisions in the decision tree independently. This allows the neural network to predict the scale of the outputs more accurately, without requiring it to also predict the relevant output probability scalings.

3.6. *Training and Test Data Split*

The training and validation data were split randomly in an 80-20 configuration, resulting in a training set size of 49,263 images, and a validation set size of 12, 315 images. Both datasets were subject to identical preprocessing before being used for training and validation, and label distributions were inspected to ensure no systematic bias existed between them in the data separation process.

3.7. *Data Generators*

Due to the large size of the dataset, the images had to be loaded in with generators in batches. The generators encapsulated functionality for rotating the dataset and producing the relevant labels corresponding to the training data. The labels and training data were thoroughly checked to match, and ensure no shuffling of the results occurred. Data was ensured to match the training labels accordingly. Additionally, the validation set was fed from a different generator instance, ensuring the datasets were not combined. At the end of each epoch of training, the generators reshuffled the training dataset, ensuring a random sample of images were taken and the model was not trained on a specific subset of the entire dataset. The generators were optimized for performance with native TensorFlow functions.

4. RESNET50 TRAINED MODEL

4.1. *Implementation*

We use an implementation of ResNet50 native in `TensorFlow` with ImageNet-initialized weights. We wrap the output layer with a Flatten layer followed by a fully-connected dense layer with 37 nodes, serving as the regression outputs. These outputs are then fed through a sigmoid and "re-normalization" custom activation function implementing the rescaling of weights as described in a previous section.

4.2. *Training Hyperparameters*

Significant testing was done to analyze different model training hyperparameters. An Adam optimizer was used, with an initialized learning rate of 0.001, and root mean squared error was used as the metric and loss function for training. A learning rate scheduler was used which dropped the learning rate by a factor of 2 after a plateau of 2

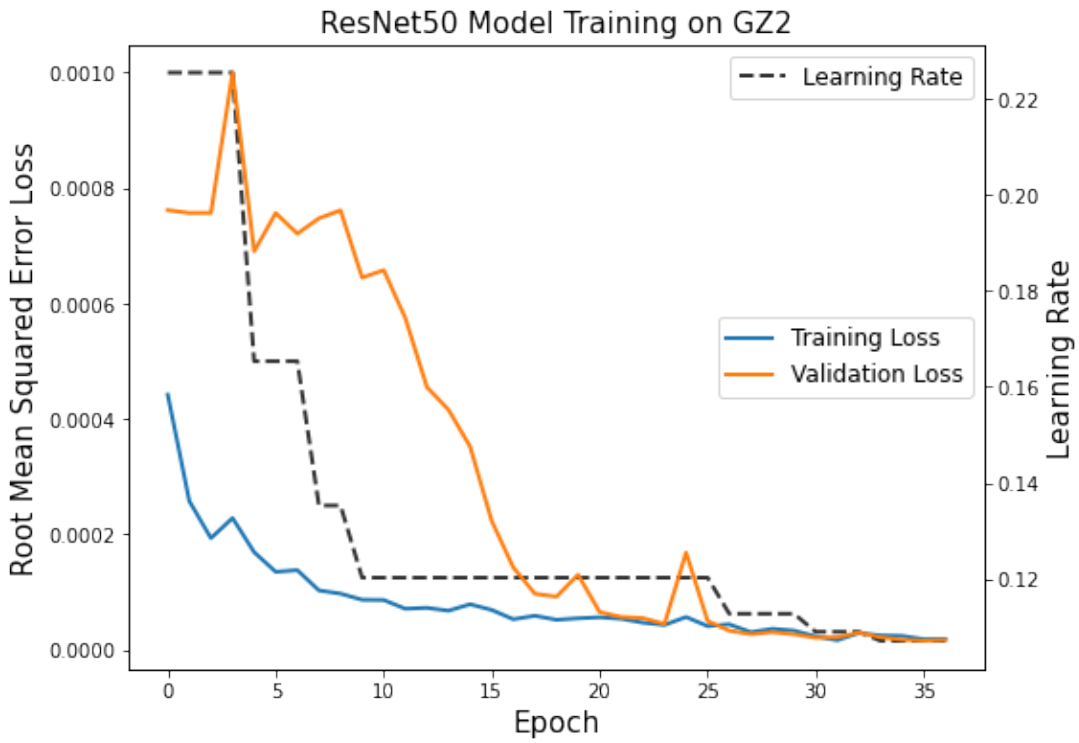

Figure 6: Loss curves and learning rate schedule for the best and longest training run, cut short by computational and time constraints. The most significant drop in validation loss peaked at a learning rate of 1.25e-4, indicating the initial model learning rate of 1e-3 may be too high. We achieve a final validation root mean squared error of 0.107, and a final root mean squared training error of 0.108. The agreement between the training and validation loss is confirmation that data augmentation was correctly implemented, disallowing consistent overfitting.

epochs of loss in training upto a minimum learning rate of 10^{-6}. Due to computational constraints, a batch size of 75 was selected, and 50 batches were trained per epoch, before the dataset was reshuffled. Training was pursued for a total of 36 epochs, in total 1,800 batches, consisting of 135,000 images with different rotational augmentations. A heavy model, the augmentation was crucial to ensure the model did not overfit to the data.

4.3. *Model Training*

Due to the low batch size and batch steps per epoch, the training loss fell much faster than the validation loss, as the model quickly overfit to the initial batches. Following which, the validation loss gradually dropped over multiple epochs and learning rate adjustments. The model trained for 3.5 hours, at a rate of 6 minutes per epoch of 50 batches, and could be trained further, with a less aggressive learning rate schedule. Due to computational and time constraints, we stop the training here. We achieve a final validation root mean squared error of 0.107, and a final root mean squared training error of 0.108.

The reason for the eventual equivalence of the validation and training error is due to the fact that after 36 epochs, the training model had most likely experienced the entirety of the training dataset being fed to it in random batches. As such, the training loss was an accurate estimate of the typical model performance, no longer fitting to any specific batch. Since the data was augmented, both the training and validation set had equal prediction quality, since on training a new batch, due to the randomized rotations, it had not seen that instance of data before in training. In other words, the training dataset carried no features that uniquely distinguished it from the validation dataset, ensuring that any model loss was a true regression loss.

We present the evolution of training loss, validation loss, and learning rate for our best model in Figure 6.

4.4. *Model Classification Quality*

The classification quality of the validation dataset can be judged by comparing the predicted and true labels. We present them in Figure 7. On training with an aggressive plateau-based learning rate schedule and just 36 epochs, the model is still capable of achieving strong results with data augmentation, output rescaling, and ImageNet initialization.

The regression results offer insight into the biggest causes of RMSE loss. Namely, multiple parameters had a large number of 0 valued labels, which the model did not train to understand, instead generally guessing a value for those parameters. This can be improved by building the decision structure into the neural network, building a different network for each decision step made. Then, if the galaxy is not odd, the model will accurately report 0 for dust lanes and lens arcs. For general classification tasks, the model performs remarkably well, classifying smooth galaxies from disk galaxies, stars from smooth galaxies, edge-on disks, bars, spiral arms, and bulge features for a spiral galaxy. The model was even able to identify odd features, and judge the shape of the elliptical galaxy: "round, cigar, or in between." The model primarily struggled with parameters with a large number of null values.

4.5. *A Note on Merger Galaxies*

The model was inspected for merger galaxies, and it was found that on merger rate predictions of 0.4 or higher, the majority of galaxies classified tend to be merging galaxies. For SDSS galaxies, such a machine learning model may be used to flag merger galaxies with a relatively high degree of certainty. However, it is clear that the merger classification makes the same error as human classifiers in certain cases, identifying irregular galaxies or galaxy doubles as merging galaxies.

5. CONCLUSION AND FURTHER WORK

With a ResNet50 model trained in `TensorFlow Keras`, we achieve a strong model prediction, with a training loss of 0.108 and validation loss of 0.107. The model accurately distinguishes important classes, namely disk galaxies, elliptical galaxies or irregular galaxies, and second-order features, like elliptical inclination, bulge structure, spiral arms, and the determination of the prescence of odd features. The model is capable of detecting galaxy mergers with caveats that it has not yet achieved superhuman classification abilities, and remains inaccurate compared to classifications by amateur or professional astronomers.

Future work remains to improve the model, including hyperparameter optimization with regards to the learning rate scheduler. An aggressive learning rate schedule was chosen, with a decay factor of 0.5 with a patience of 2 epochs. Due to random stochasticity, a close-to optimal learning rate may be skipped in favor of a lower learning rate. Near the end of the model training, this becomes clear, with the learning rate dropping to very low values and effectively arresting the dropping loss of the model. With greater time and computational resources, a more patient learning rate schedule and greater number of epochs can gain the maximum training from the augmented data and may produce improved classification quality.

REFERENCES

Conselice, C. J. 2014, ARA&A, 52, 291, doi: 10.1146/annurev-astro-081913-040037

Lintott, C. J., Schawinski, K., Slosar, A., et al. 2008, MNRAS, 389, 1179, doi: 10.1111/j.1365-2966.2008.13689.x

Metcalf, R. B., Meneghetti, M., Avestruz, C., et al. 2019, A&A, 625, A119, doi: 10.1051/0004-6361/201832797

Sreejith, S., Pereverzyev, Sergiy, J., Kelvin, L. S., et al. 2018, MNRAS, 474, 5232, doi: 10.1093/mnras/stx2976

Willett, K. W., Lintott, C. J., Bamford, S. P., et al. 2013, MNRAS, 435, 2835, doi: 10.1093/mnras/stt1458

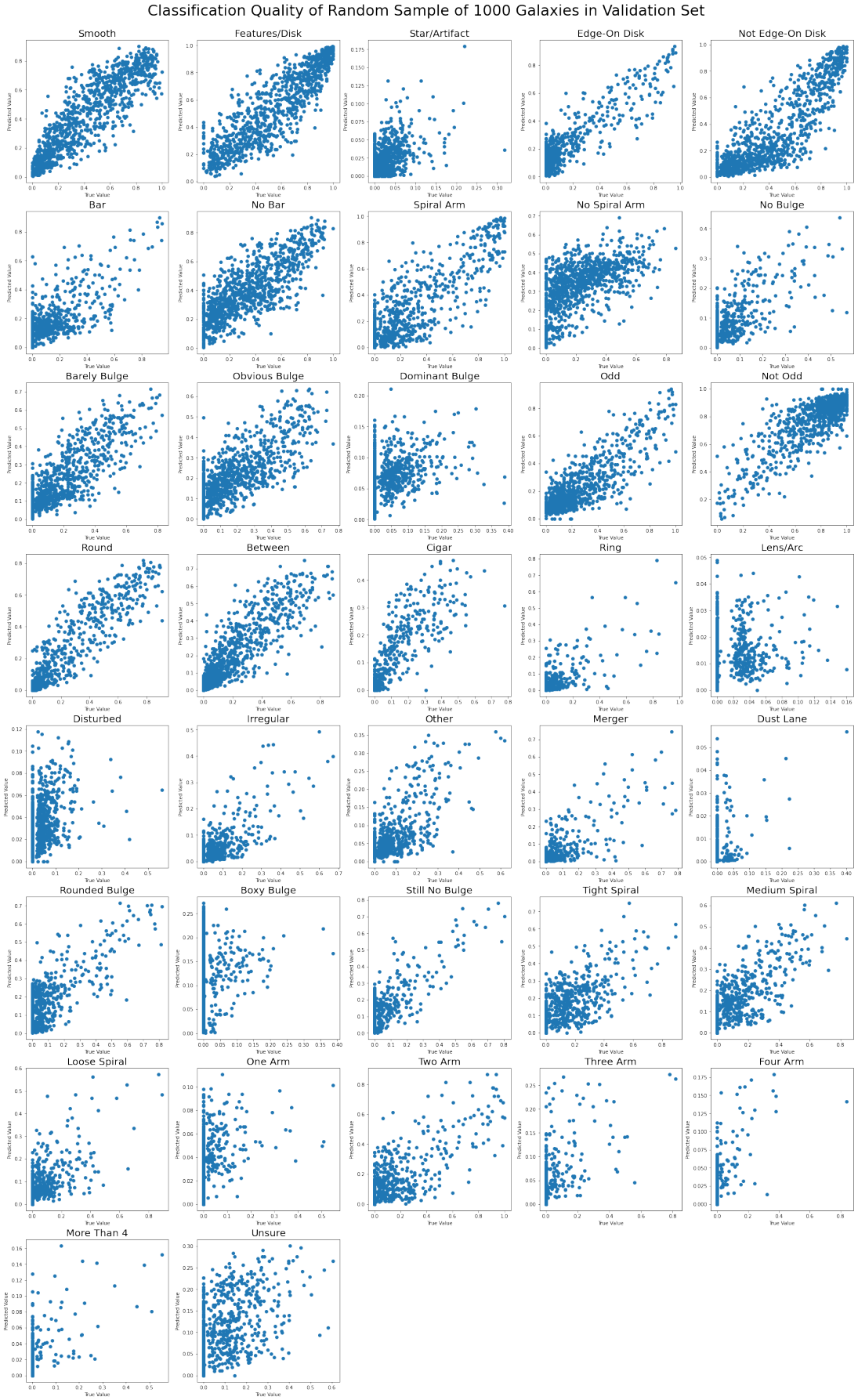

Figure 7: Model prediction against true data for 1000 randomly sampled galaxies in the validation set. True labels are presented on the x-axis, while predicted labels are on the y-axis. We find that the model accurately predicts a smooth or disk galaxy, as well as multiple other parameters.