UNIVERSITY OF CALGARY

DATA 608

GROUP PROJECT FINAL REPORT

Classification of Galaxy Zoo 2 Images

Ryan JOHNSTON UCID: 10172915
Scott PELLEGRINO UCID: 10150458
Troy STOLZ UCID: 10058345

April 2, 2021

Contents

List of Figures

List of Tables

Introduction and Background

Contrary to what one might think, galaxies are a relatively new concept in astrophysics. What we now call galaxies were once thought to be “spiral nebulae” that existed inside the Milky Way, and it was believed that the Milky Way was the entirety of the universe (Curtis 1917).

This eventually became what was known as "The Great Debate" in astronomy between Harlow Shapley and Heber Curtis (Shapley & Curtis 1921). Harlow Shapley argued that spiral nebulae such as Andromeda were simply part of the Milky Way. If the nebulae were extragalactic the distance to it must be $\sim 10^7$ parsecs (pc), which he thought would be a ridiculously large distance, something most astronomers at the time agreed with (1 pc is 3.26 light-years). However, Heber Curtis argued that the nebula were in fact “island universes” and existed outside of our galaxy. He based his argument on the fact that the known number of novae in a small angular section of Andromeda would imply a huge Nova rate in a small physical region of the sky. In other words, if Andromeda were part of the Milky Way there would have to be more novae in one small section of the galaxy than the rest of the Milky Way combined. This was a very hard point to explain unless Andromeda existed far outside outside our own galaxy.

In the 1920s, Edwin Hubble resolved the debate with his observations of variable stars in galaxies (Hubble 1929). Cepheid variable stars are one of the earliest methods that astronomers used to determine the distance to other galaxies. This is due to the fact that there is a direct relationship between the star’s luminosity and its pulsation period. Observing the pulsation period allows one to derive the true luminosity of the star, and from there you can determine the distance by comparing it to its observed brightness. Edwin Hubble determined that the distance to Andromeda to be at least 300 kpc away, compared to the modern value of 800 kpc. By comparison, the Milky Way is only $\sim$ 30 kpc in diameter. It was only then that astronomers finally realized that galaxies were an entirely new class of objects that were far beyond the Milky Way. According to Conselice et al. (2016), there are an estimated two trillion galaxies in the observable universe.

During this same time period, modern physics was still under development. For example, Einstein’s derivation of the mass-energy equivalence $E = mc^2$ was published in 1905 (Einstein 1905), and it wasn’t until May of 1919 that the first observational test of Einstein’s theory of General Relativity, gravitational lensing, was confirmed during a solar eclipse (Dyson et al. 1920).

The Great Debate took place one century ago on April 26, 1920 at the Smithsonian Institute. Today, the formation and evolution of galaxies is still one of the major topics of research in astrophysics Many questions still remain about the vast array of galaxy morphology’s, properties, and processes that exist. These include the way galaxies first formed in the universe, how galaxies evolve and interact with their local environment, and what the eventual fate of galaxies are.

In this project we aim to utilize various machine learning techniques to classify galaxies based on their morphology. We will use classical machine learning techniques such as naive Bayes and random forest classifiers as well as deep learning techniques such as central neural networks. We also aim to utilize the Teaching and Learning Cluster (TALC) to help with processing, model training, and prediction.

Galaxies in the Universe

Galaxy morphology is important because it is strongly correlated with other physical parameters. These include the surface brightness of the galaxy and being a tracer of the orbital mechanics of the stars inside it. It is also an indicator of star formation. An example of how galaxies are classified is shown in Figure 1, with some actual image examples shown in Figure 2.

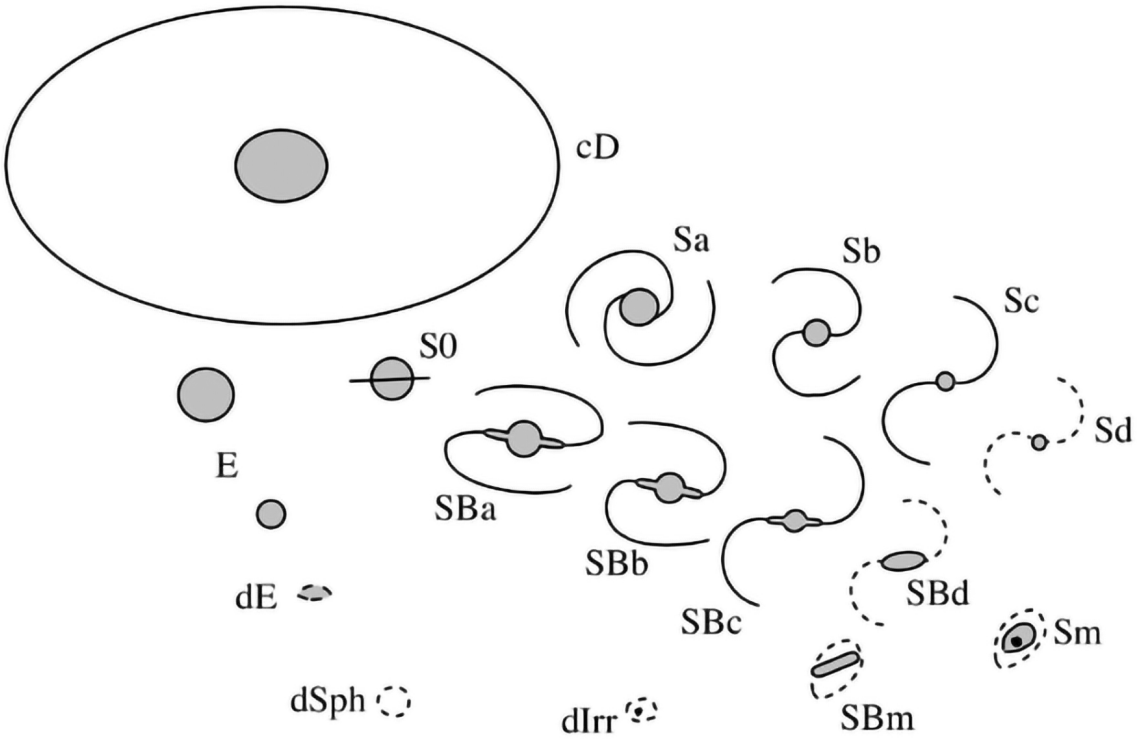

Fig 1.11 'Galaxies in the Universe' Sparke/Gallagher CUP 2007

Figure 1: A Galaxy classification system taken from Figure 1.11 from Sparke (2007) which follows a modified version of the Hubble tuning fork diagram. It adds key elements from the Yerkes classification scheme (Morgan & Osterbrock 1969). It is also much simpler than the de Vaucouleurs-Hubble scheme (de Vaucouleurs 1959).

In general the classification scheme of galaxies goes as follows:

- cD or giant elliptical galaxies.
- E –Elliptical galaxies.
- S0 - Lenticualr galaxies.
- S – Spiral galaxies.
- SB – Barred Spiral galaxies.

- Irr – Irregular galaxies.
- d – Dwarf galaxies.

Giant elliptical galaxies, also known as cD galaxies, denotes a massive central galaxy in galaxy clusters. They have a dense core and an extended halo and can be > 100 times more luminous than the Milky way. Elliptical galaxies (E) are shaped like a spheroid and smooth, and appear almost featureless. Spiral galaxies are disk shaped galaxies that have a central bulge or bar with bright spiral arms. About 50% of spirals have central bars (SB galaxies) and 50% do not (S galaxies). As we progress from SBa to SBc and S to Sc the central bulge becomes less prominent compared to the disk and the arms become more open. Eventually these arms become more disorganized (Sd, SBd). Finally, we have Lenticular galaxies (S0) as the last major category of galaxy. Lenticular galaxies have the disk shape and bulge of a spiral galaxy but have no bar or spiral arms present. These seem to represent a transition between ellipticals and spirals. Everything else can be classified as irregular galaxies.

In this project we opted to keep a simple binary classification of that galaxies between elliptical galaxies, also known as 'early type' galaxies, and spiral galaxies, also known as 'late type' galaxies. Spiral galaxies are thought to contain ~ 70% of the galaxy population, and conversely elliptical galaxies are thought to occupy ~ 30% of the population (Buta et al. 2015).

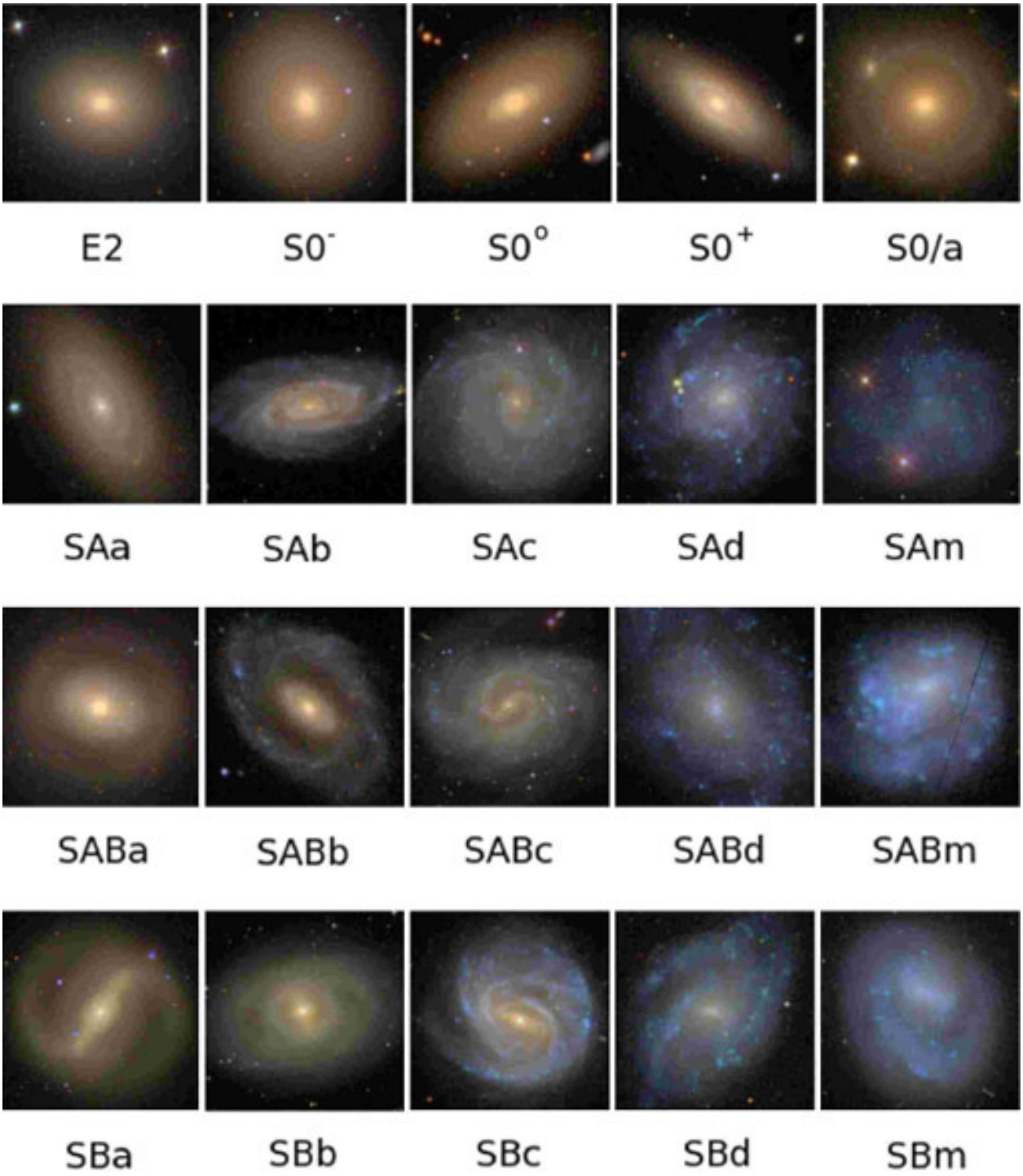

Figure 2: SDSS examples of galaxy classification using the Hubble tuning fork (Hubble 1926). This image mosaic is taken from Figure 48 of Buta (2011).

The Data

The Sloan Digital Sky Survey

The Sloan Digital Sky Survey (SDSS) is a major multi-spectral imaging and spectroscopic redshift survey (Gunn et al. 2006, York et al. 2000, Eisenstein et al. 2011, Blanton et al. 2017). Data collection first began in 2000 and to date covers over 35% of the sky, with photometric observations of around nearly 1 billion objects with additional spectrosopic data continually being taken. It is often hailed as the most ambitious astronomical survey ever undertaken. The first results of the next stage of the survey wil published this July. The data we will be using in this project is from the SDSS Data release 8 (DR7) and 8 (DR8).

The SDSS uses five filters described in Table 1 and shown graphically in Figure 3.

Filter	Average Wavelength (nm)
Ultraviolet (u)	354.3
Green (g)	477.0
Red (r)	623.1
Near Infrared (i)	762.5
Infrared (z)	913.4

Table 1: Filters used by the imaging camera of the SDSS.

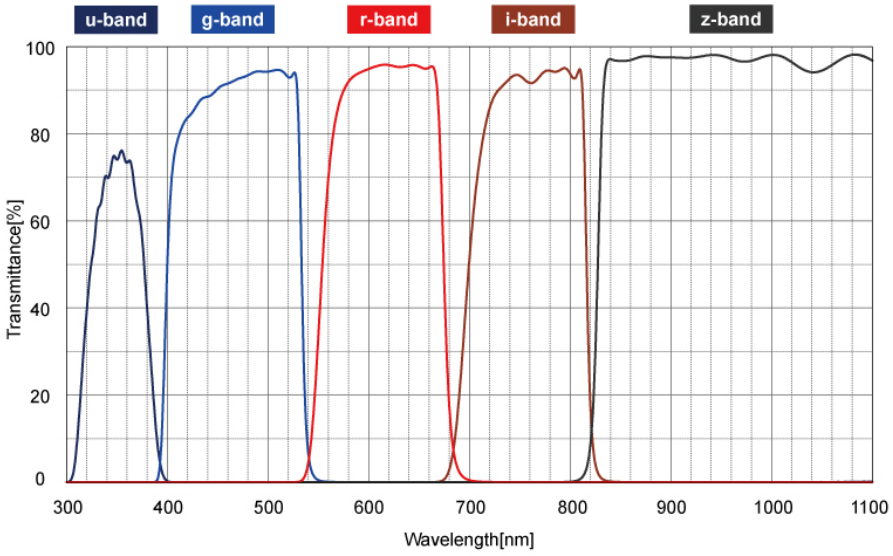

Figure 3: Plot of the transmittance of the SDSS filters over wavelength. Source: https://www.asahi-spectra.com/opticalfilters/sdss.html

The Galaxy Zoo Project

In this project, we make use of the second iteration of the Galaxy Zoo project (GZ2). The galaxy zoo projects relied on "citizen scientists" and other human volunteers to classify nearly 300,000 galaxies in the SDSS (Willett et al. 2013, Hart et al. 2016). The citizen scientists are asked a specific set of multiple choice questions about the galaxy in the image and depending on the outcome would determine the class of galaxy. If enough people who view the image come to a consensus then that is likely the class the galaxy actually is. The scientists in charge of the galaxy zoo project go into detail about the steps they took to de-bias the classification results (Hart et al. 2016). Beside the images, we are also interested in the galaxy classifications and the metadata table, which provides various scientific information about the selected galaxies and their classifications. Therefore, it is assumed in this project that all of the classifications provided are true as the tables are de-baised beforehand.

The images from the galaxy Zoo Project are available for downloaded from Kaggle with a supplmentary table on objectID and file name matching. There are 239,571 composite JPEG images made from the g, r, and i SDSS filter bands. They occupy about 3 Gb on disk when uncompressed. The pixels in each SDSS image is 24μm in size and has a pixel scale of 0.396arcsec/pixel. Meaning that each 212 by 212 image captures 17797.824 arcsec^2 or $\approx$ 4.944 degrees^2 of the sky (1 degree = 3600 arsec). An example of these images are shown in Figure 5.

For the purpose of this project we cross reference with three other tables to provide us with additional information. We use the table from the first iteration of the Galaxy Zoo (GZ1), it provides us with a simple broader classifications (Lintott et al. 2008, 2011). In GZ1, there are 122,134 images that contain 'uncertain' galaxies. Meaning that not enough people could agree on what kind of galaxy it was, or too few people attempted to classify the galaxy.

Additionally, we utilize a table from the GZ1 that describes galaxies that host an active galactic nucleus (AGN) (Schawinski et al. 2010). These AGNs are powered by super-massive black holes at the centre of the galaxies. The table also includes starburst galaxies. Another table associated with the GZ1 catalogue discusses merging galaxies detected in the Galaxy Zoo (Darg et al. 2010). All of these tables can be found at the Galaxy Zoo web-page. The tables occupy about 400 mb on disk in total.

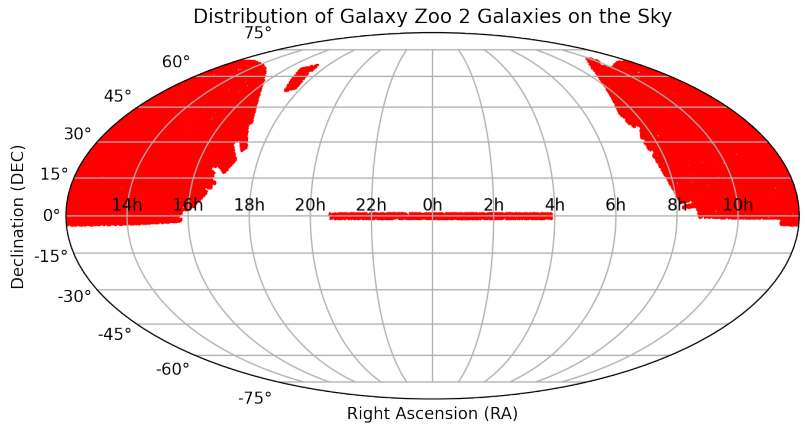

Figure 4: A Molliweide projection of the galaxies in the GZ2 catalogue appear on the night sky.

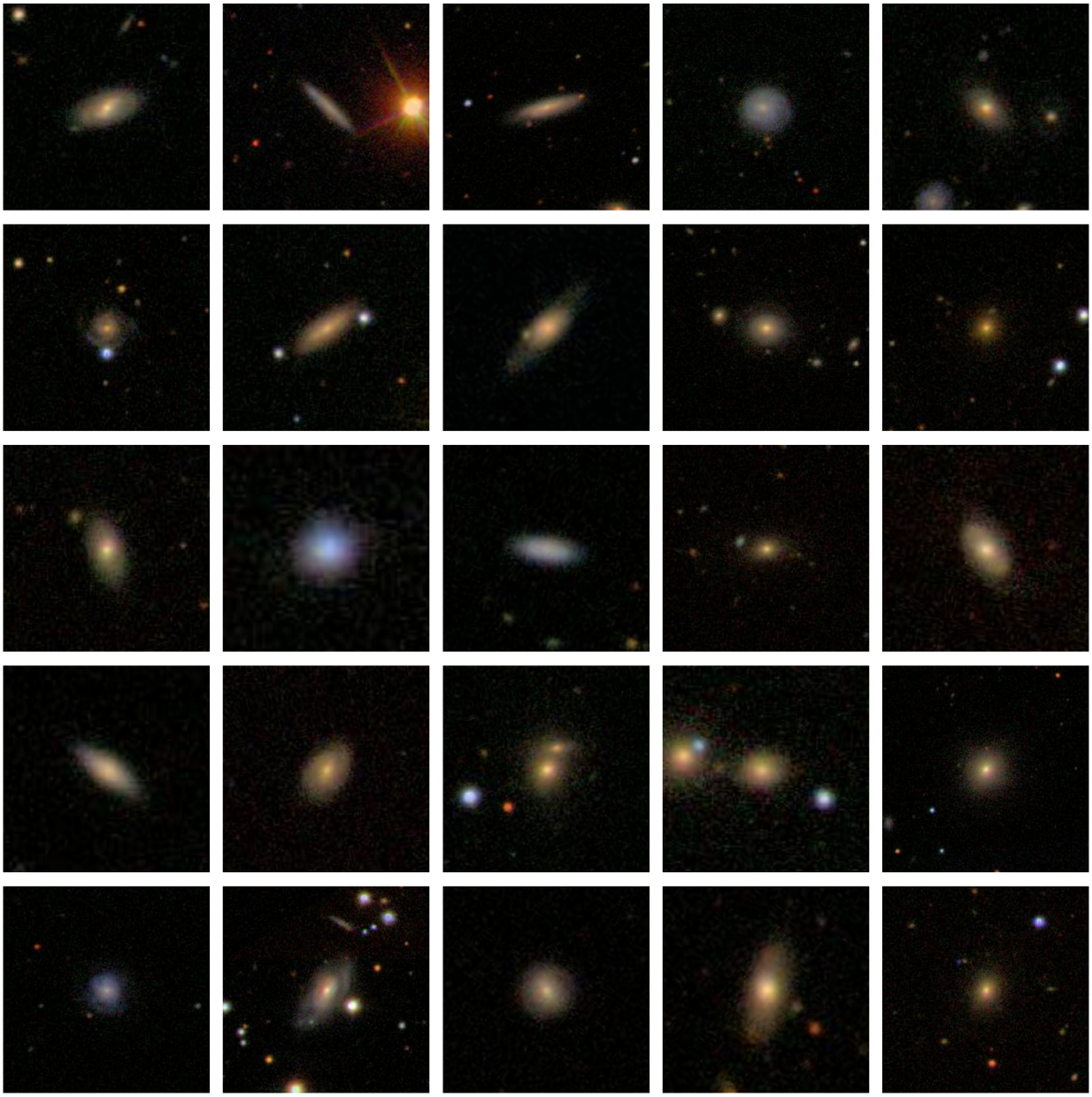

Figure 5: A random sample of 25 galaxies from the data. The images are in their original resolution of 212 × 212 pixels before any image processing is done on them. Panel (1,2) has a star in the foreground. You can tell because of the spokes present in the image.

Methodology

Data Cleaning and Merging

The *Galaxy Zoo 2* table, *Galaxy Zoo 1* table, *Galaxy Zoo: Mergers*, and *Galaxy Zoo: AGNs* tables were merged using the `Pandas` Python library. Each of the tables was easily merged into one table using the SDSS DR7 `OBJID` (Object ID) as the common key. The images obtained from

Kaggle also include a CSV file which maps the object ID to the image file name, so that table was merged with the other as well. Other appropriate cleaning and wrangling was done as needed, but this was minimal. Columns that were not of interest were dropped from the data frame and exported as a CSV for further use in our model construction.

The final combined table has the following variables:

- **OBJID** – Identifier of the object.
- **ASSEST_ID** – JPEG file extension of the image associated with the galaxy.
- **RA** – Right Ascension of the Galaxy in degrees.
- **DEC** – Declination of the Galaxy in degrees.
- **REDSHIFT** – The spectroscopic redshift of the galaxy (how far away it is).
- **gz2_class** – The morphology of the galaxy according to Galaxy Zoo 2.
- **GZ1_MORPHOLOGY** – The morphology of the galaxy according to Galaxy Zoo 1.
- **STAR_FORMING** – Flag if the galaxy is undergoing a star forming burst or not.
- **AGN_LINER** – Flag if the galaxy hosts an AGN or LINER.
- **PETROR50_R** – The 50% Petrosian angular radius in the r-band in arcseconds.
- **PETROR90_R** – 90% Petrosian angular radius in the r-band in arcseconds.
- **PETROMAG_U** – Petrosian apparent magnitude in the u-band.
- **PETROMAG_G** – Petrosian apparent magnitude in the g-band.
- **PETROMAG_R** – Petrosian apparent magnitude in the r-band.
- **PETROMAG_I** – Petrosian apparent magnitude in the i-band.
- **PETROMAG_Z** – Petrosian apparent magnitude in the z-band.
- **PSFMAG_R** – Point spread Function (PSF) magnitude in the r-band.
- **FIBERMAG_R** – Spectroscopic fiber (3 arcsecond) magnitude in r-band.
- **DEVMAG_R** – Model magnitude from a pure deVaucouleurs profile fit in the r-band.
- **EXPMAG_R** – Model magnitude from a pure exponential profile fit in the r-band.
- **FRACDEV_R** – Coefficient of deVaoucouleurs component in linear combination of models in the r-band.
- **MU50_R** – The surface brightness within the 50% petrosian r-band radius.
- **CMODELMAG_R** – The composite model magnitude in the r-band.
- **PETROMAG_MU** – The petrosian absolute magnitude in the u-band.
- **PETROMAG_MG** – The petrosian absolute magnitude in the g-band.

- **PETROMAG_MR** – The petrosian absolute magnitude in the r-band.
- **PETROMAG_MI** – The petrosian absolute magnitude in the i-band.
- **PETROMAG_MZ** – The petrosian absolute magnitude in the z-band.
- **PETROR50_R_KPC** – The physical Petrosian half-light radius in kilo-parsecs (kpc).
- **PETROR50_R_KPC_SIMPL_BIN** – The bin number for the physical size of the galaxy.
- **PETROMAG_MR_SIMPLE_BIN** – The bin number for the luminosity of the galaxy.
- **REDSHIFT_SIMPLE_BIN** – The bin number for the redshift of the galaxy.

Tools

In this project we make use of `scikit-learn`, `tensorlfow`, `keras`, the `Python Imaging Library` (PIL) and, `multiprocessing`. We also use the common data science libraries such as `Pandas`, `Matplotlib`, and `Seaborn`. Finally, we used the Python library `astropy`[1] to read in FITS tables and convert them to pandas dataframes for merging. We also used `astropy` to help with coordinate transformations for the creation of Figure 4.

Through some testing, we found that with multiprocessing and GPU usage, computation times and general efficiencies were more favourable on our local machines, rather than using TALC or Google Colabs. Therefore, wherever possible multiprocessing and GPU acceleration were used.

Random Forest

Variables of Interest

The main features of interest that we will be using to do our galaxy classification using classical machine learning techniques are color index, adaptive moments, eccentricities and concentrations. All of these variables are either part of the GZ2 metadata table or can be computed from it.

Color indices are simply the difference between two magnitude measurements of the galaxy at different wavelengths. The magnitudes found at the longer wavelength being subtracted from that at the shorter wavelength. In the SDSS this is u – g, g – r, r – i, and i – z. Colour indices are useful because spiral galaxies tend to have younger star populations and are therefore brighter at lower wavelengths (i.e. – 'bluer'). Conversely, elliptical galaxies tend to have an older star population and are thus brighter at higher wavelengths (i.e. – 'redder').

We will we using three other features to describe the shape of the galaxy. Eccentricity approximates the shape of the galaxy by fitting an ellipse to its profile. Eccentricity is the ratio of the two axis (semi-major and semi-minor). The De Vaucouleurs model was used to attain these two axis. The De Vaucouleurs model is defined as

$$I(R) = I_e \exp\left(-7.669\left[\left(\frac{R}{R_e}\right)^{1/4} - 1\right]\right),$$

[1] https://www.astropy.org/

where R_e is the effective radius where the brightness decreases by half its central value and $I_e = I(R_e)$ is the intensity of the corresponding isophote (de Vaucouleurs 1948) or on the SDSS technical website.

Adaptive moments area another variable that also describes the shape of a galaxy, they are the second moments of the object intensity. We use the fourth-order moment here for each filter band. Defined as

$$m_{cr4} = \frac{\langle r^4 \rangle}{\sigma^4},$$

where m is the magnitude, r is the radius, and σ is the size of the gaussian weight. They are used in image analysis to detect similar objects at different sizes and orientations. More details on adaptive moments can be found in Bernstein & Jarvis (2002).

Concentration is similar to the luminosity profile of the galaxy, which measures what proportion of a galaxy's total light is emitted within a certain radius. A simplified way to represent this is to take the ratio of the radii containing 50% and 90% of the Petrosian flux.

The Petrosian method allows us to compare the radial profiles of galaxies at different distances. More information on the Petrosian approach is available here or here. We will use the concentration from the u, r and z bands. The concentration can be defined as folows,

$$\text{Conc} = \frac{\text{Petro}_{\text{R50,a}}}{\text{Petro}_{\text{R90,a}}}$$

where a is one of the SDSS filter bands.

All of these features are found as part of the SDSS photometric catalogue, and are provided in the metadata table for GZ2.

Multinomial Naive Bayes Classifier

The first model constructed to classify the galaxy data was a Multinomial Naive Bayes. The Naive Bayes method is based off of Bayes Theorem which describes the probability of an event to occur based on prior knowledge of conditions which may be related to the event. It is determined using the following equation.

$$P(A|B) = \frac{P(A) * P(B|A)}{P(B)}$$

For our purposes, the event we are interested in determining is the classification of the galaxy, and the events which we have a prior knowledge of are the galaxy parameters collected in the Galaxy Zoo 2 metadata.

To produce a Naive Bayes model using Bayes theorem, we must assume that all parameters are independent of one another. Additionally, as multiple parameters are being used in the classification, the model becomes multinomial, hence a Multinomial Naive Bayes classifier.

A limitation of the Multinomail Naive Bayes classifier is that it cannot handle negative values. As such parameters within the galaxy data which had both positive and negative values were omitted and the absolute value of strictly negative parameters was taken.

Random Forest

To further utilize the numerical data, a random forest classifier was trained. The random forest classifier constructs multiple decision tree classifiers and reports the majority decision. For each constructed tree, the points where it branches, or decisions are made, will differ slightly. While training, the model will adjust the decision points and values to minimize the sum of square error (SSE).

While it is possible to construct random forest classifiers using image data, for this investigation, only the measurements collected from the supplemental datatsets will be used.

Convolutional Neural Network

Another model that we will attempt to construct is a convolutional neural network (CNN) using the tensorflow and keras libraries. A neural network is a model that takes inspiration from the brain. It is composed of layers of which at least one is hidden. It consists of simple connected units (or neurons) followed by non linearities. This means that a convolutional neural network is a neural network in which at least one layer is a convolutional layer. A typical convolutional neural network consists of some combination of convolutional layers, pooling layers, and dense layers. These types of neural networks tend to have great success in image recognition problems such as the one we are trying to solve.

A convolutional operation follows a two step mathematical approach:

1. Element-wise multiplication of the convolutional filter and a slice of an input matrix.
2. The summation of all the values in the resulting product matrix.

Convolutional layers consist of a series of convolutional operations, each acting on a different slice of the input matrix.

A pooling operation is typically performed after a convolution. This is done by again dividing the matrix into slices and then slides the convolutional operation by strides. For our purposes we will be using max pooling which selects the largest value out of the slices.

A dense layer is a synonym for a fully connected layer. This is the hidden layer in which each node is connected to every node in the subsequent hidden layer. Since we are trying to predict six different categories our final dense layer will be of size six.

Results

Naive Bayes

Using the numerical data compiled from the Galaxy Zoo 1 and Galaxy Zoo 2 projects, a multinomial Naive Bayes classifier was trained. Of the available 115,225 entries, 80% of these were allocated as training points, with the remaining 20% reserved for testing. As stated previously, the Naive Bayes implementation cannot accept negative values. A simple solution or workaround to this would be to take the absolute value of all parameters. Unfortunately, there are some parameters with both positive and negative values. If the absolute value were to be taken of these parameters, the meaningfulness of the value would be lost. For this reason, parameters with both positive and negative values were excluded when creating the Mulitnomial Naive Bayes model. These parameters were found to be "PETROMAG_MU", "PETROMAG_MI", "PETROMAG_MZ", and "PETROMAG_MR_SIMPLE_BIN". As removing these parameters does not remove all negative values, the absolute values of any remaining negative values was taken. With these incompatible parameters removed, 24 usable parameters remain.

Using the MultinomailNB function from the SciKitLearn package, a multinomial Naive Bayes model was constructed, using the training data. This model produced a classification accuracy of 74.9% when classifying the testing data. To determine if model over fitting was present, the model was given back the training data to predict. A training classification accuracy of 75.5% was observed. As this value is relatively similar to the testing accuracy, over fitting is not believed to be significantly influencing prediction results.

This model is the most simplistic that we tested and, as will be seen, produced the lowest accuracy values. Due to the accuracy of roughly 75%, depending on the need, this model would likely not be suitable for reliable prediction of unknown galaxies.

Random Forest

Following the Multinomial Naive Bayes implementation, a random forest classifier was created. This model was created using the same data as the Naive Bayes classifier, with the exception that no parameters were removed and the sign of values was left unchanged. The size of the training dataset also remained as 80% of the original 115,225 entries, and the testing dataset remained as the other 20%.

As numerous regression trees are created for a random forest, to reduce the computation times parallelization was used. Initially, these computations were intended to be completed using the Teaching and Learning Cluster (TALC), provided by the University of Calgary. Upon some testing, it was found that our local machines were able to complete the identical operations in a considerably shorter time. For example, fitting data to an initial random forest model required 4 minutes and 57 seconds to complete on the TALC network, while the same operation on our local machines took 3.36 seconds.

Using the training data, a preliminary model was produced. Using the testing data, a testing accuracy of 94.3% was found. In the same way as the Naive Bayes, the training data was fed back into the model to predict the classifications and determine if over-fitting was present. A

training error of 100% was found. As the training accuracy was notably larger than the training accuracy, over-fitting was present and required addressing.

To reduce over-fitting, parameters of the random forest (hyper parameters) were adjusted. These hyper parameters were "n_estimators", "max_depth", and "max_features". These control the number of classification trees used in a forest, the depth, or number of decisions in a tree, and the number of features to consider when deciding on the best split respectively.

Figure 6 below illustrates the random forest training and testing error for different values of n_estimators. It can be seen that values above roughly 15 begin to settle on model ac curacies. It is also worth noting that the distance between the training accuracy (black) and the testing accuracy (blue) remains roughly constant for all tested values. This means that over-fitting cannot be managed by adjusting the number of estimators.

Next, the plot investigating the effects of max_depth, shown as Figure **??**, illustrates that the testing accuracy begins to plateau around a value of 15 once again. It can also be seen that the training accuracy increases at a slower rate than the training accuracy. This can be used to reduce the level of over-fitting in the model. It appears that maximum depth values between roughly 8 and 15 can reduce the over-fitting at a minimal cost to testing accuracy.

The final hyper parameter which we investigated was max_features. From Figure 8 below, it can be seen that accuracy for both testing and training datasets do not appear to vary substantially with differing max_feature values. Additionally, there are no values which considerably reduce the distance between the two accuracies.

Finally, from testing these hyper parameters, we chose to use an n_estimator value of 20, a max_depth value of 10, and chose to use the default option for max_features, which is the number of features in the data, in this case 28.

With these hyper parameters, a newly trained random forest classifier produced a testing accuracy of 93.6% and a training accuracy of 95%. This is a decrease of 0.7% testing accuracy, and 5% training accuracy. While this model may still have minor over-fitting, as indicated by the training accuracy being larger than the testing accuracy, it is considerably reduced from the original, un-modified model. Figure 9 below is a confusion matrix communicating the final random forest model predictions and the number of each correct and incorrect prediction.

Convolutional Neural Networks

Turning now to the galaxy images, several different convolutional neural networks were created in attempts to classify galaxies based on the images alone.

Data Conditioning

Prior to training the models, some conditioning was conducted on the images to prepare them, and reduce the required computation time. The first conditioning step conducted was cropping the image. As the galaxy of interest was always in the centre of the image, the outside edges were cropped. We found that removing 25% of the image from each edge removed the majority of unnecessary information while still allowing for a small buffer between the galaxy extents

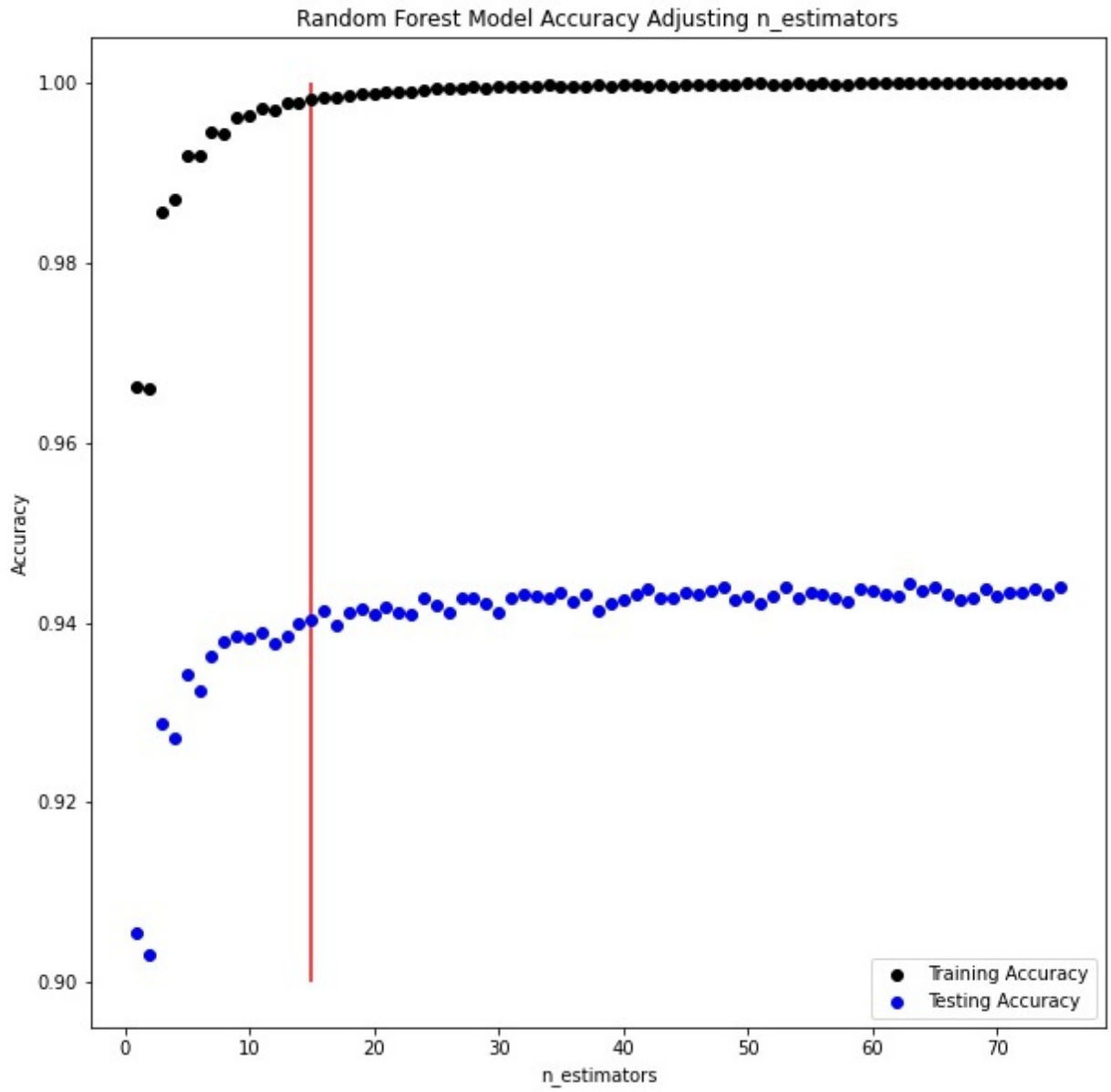

Figure 6: A plot illustrating the effect of the number of estimators on the a random forest classifier. The red line indicates the value of 15, discussed above.

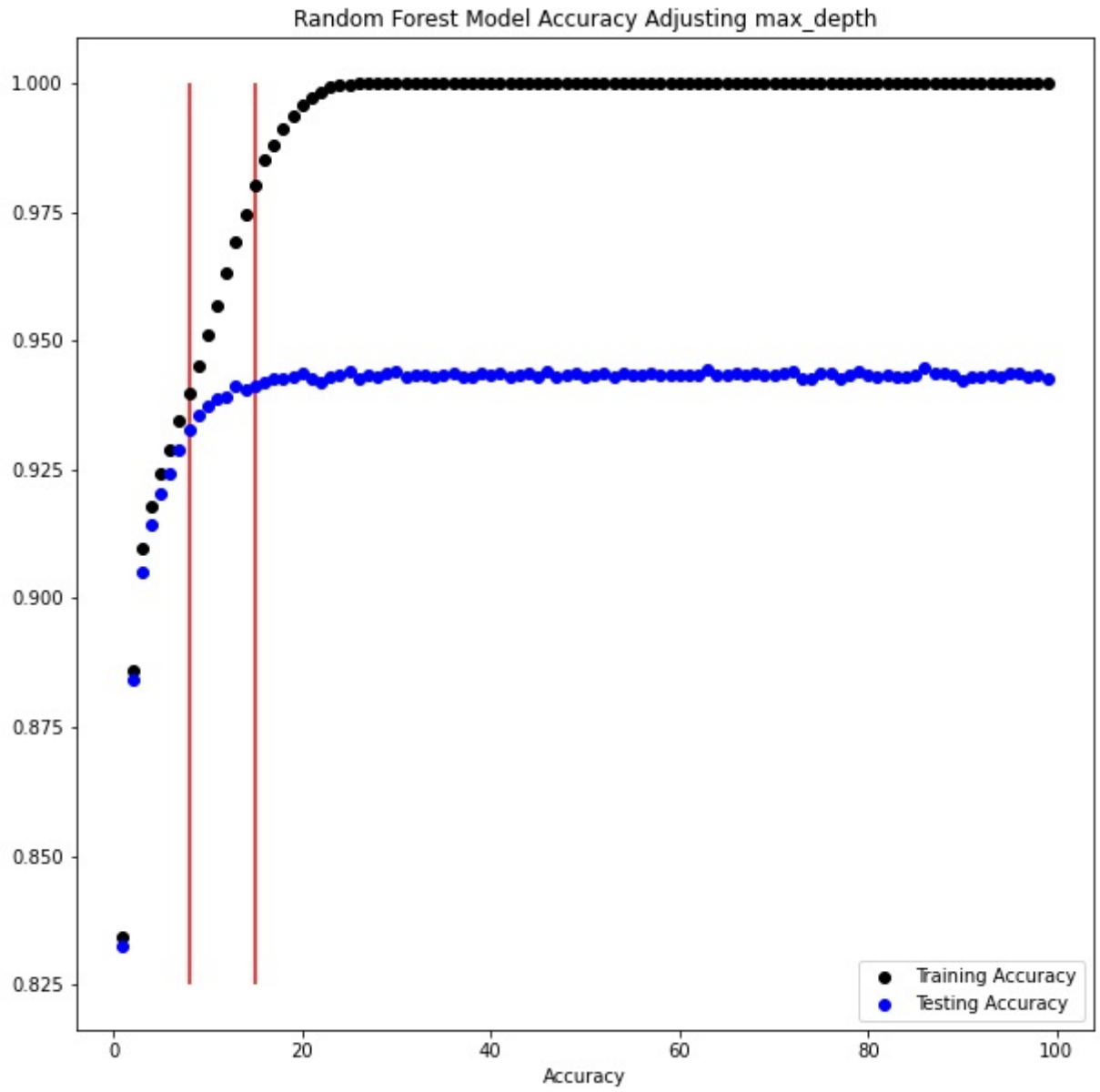

Figure 7: A plot illustrating the effect of the maximum depth on the random forest classifier. The red lines indicate the values of 8 and 15, as discussed above.

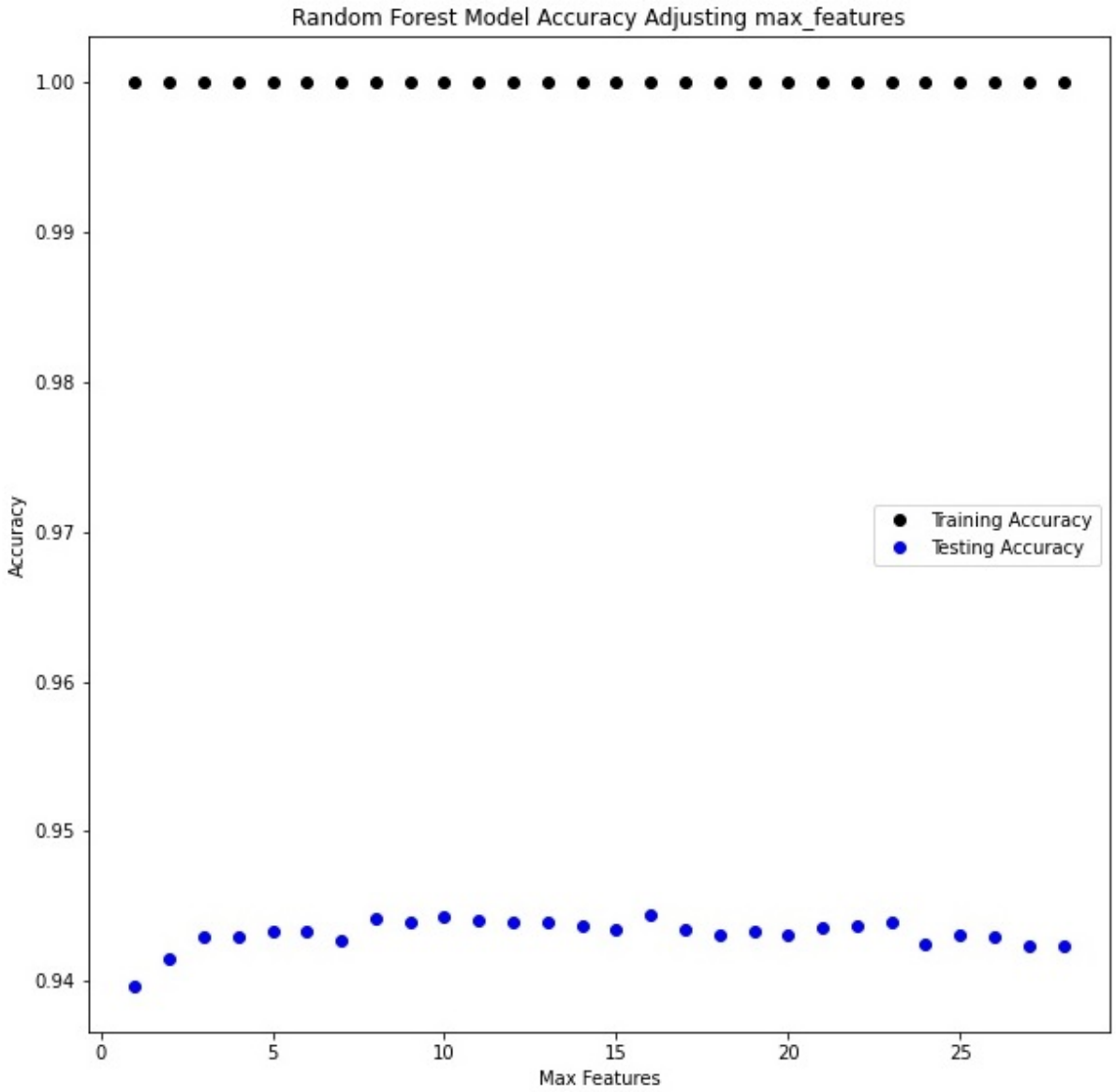

Figure 8: A plot illustrating the effect of the the maximum number of features in the random forest classifier

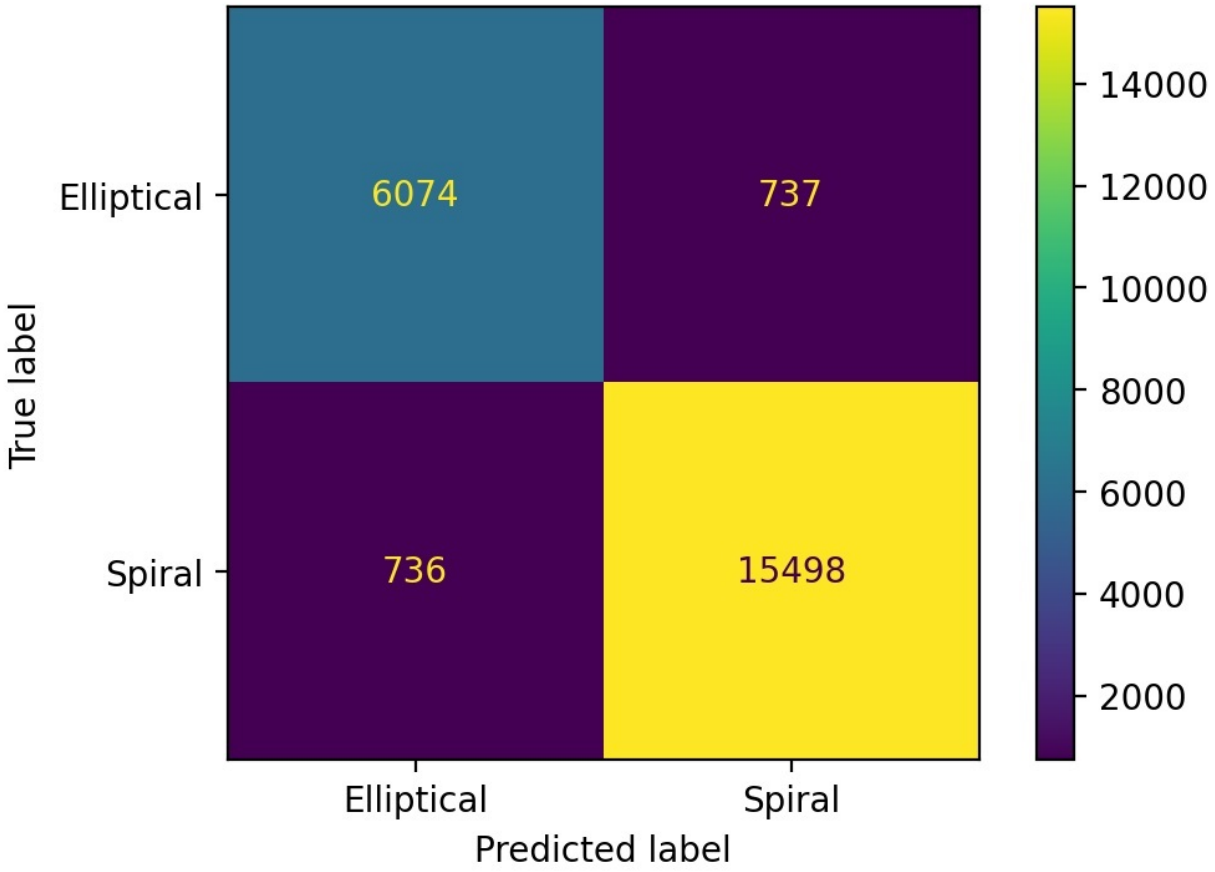

Figure 9: A confusion matrix illustrating the predictions made by the tuned random forest classifier model. The model accurately classified 6074 galaxies as elliptical and 15489 as spiral. The model incorrectly classified 736 galaxies as elliptical and 737 as spiral.

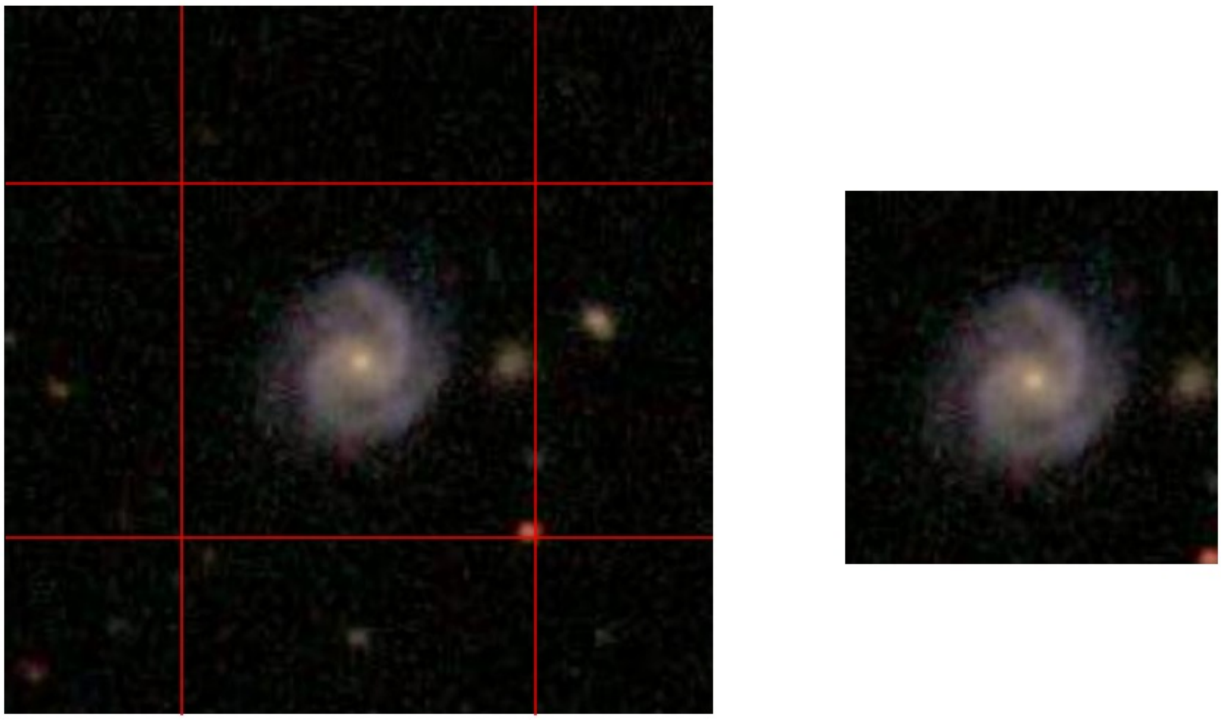

Figure 10: An example of how each image was cropped. Red lines indicate where the original image was cropped to. The image on the right is the result of the crop.

and the image bounds, in most cases. This crop reduced the dimensions of each image from 424×424 to 212×212, and reduced the storage size of the image library from roughly 3 Gb to roughly 1 Gb. Figure 10 below is an example of this process.

There were occasions were some potentially important galaxy information was removed with the crop. To determine how ubiquitous this removal was, the average luminance value for all removed bands was determined. Luminance is a measure of a pixel value intensity. There are several methods to calculate it, however we chose to use a simple equation, as shown below.

$$\text{Luminance} = 0.3\text{R} + 0.59\text{G} + 0.11\text{B}$$

The removed value was compared against the luminance values for the edges of the cropped image. If the luminance along the edges of the cropped image was more than twice the average value of the removed areas, the image was flagged. As this is a computationally intensive procedure, a random subset of 20%, roughly 48,000, of all images were checked. This included all galaxy classifications, including uncertain. Figure 11 illustrates an image where some potentially significant information was removed with the crop.

It was found that approximately 1% of these images were flagged. It is believed that due to the size of the dataset, this 1% can reasonably be extrapolated to the entire image library. This means that of the available 239,571 images, roughly 2,400 may have some amount of potentially characteristic information removed. Upon investigating some of the flagged images, it

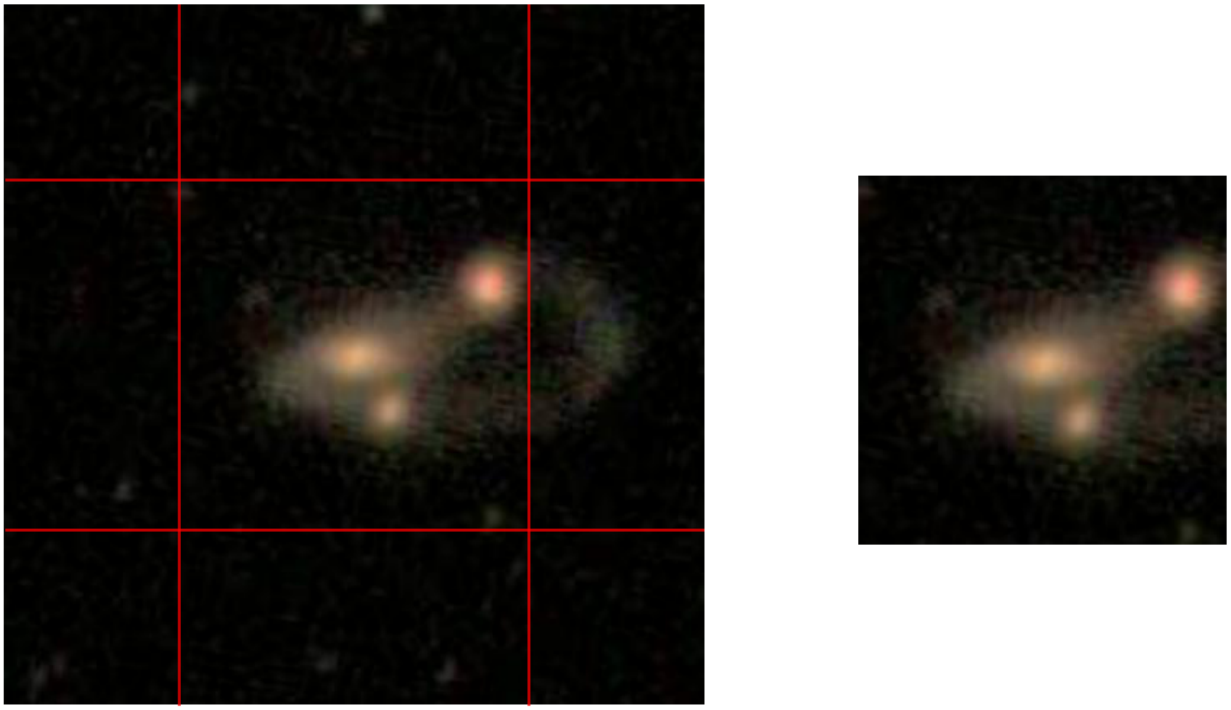

Figure 11: An example of how some galaxy information may be removed as a result of the crop.

was found that some flags were caused by ambient luminance from nearby galaxies. Additionally, the noise of some images caused the image to be flagged. Examples of both of these can be seen in Figure 12.

In attempts to remedy issue associated with noise, a simple neural network was used to reduce the image noise. This network was constructed with one fully connected hidden layer with 256 parameters and an output layer with as many parameters as pixels in the original image. For our 212x212 images, the output layer requires 44,944 parameters. To use the images with this network, they are split into the comprising red, green, and blue channels. As these are 8 bit images, the possible pixel values for each image are between 0 and 255 inclusive. It is for this reason that the hidden intermediate layer of the neural network has 256 parameters. To train this model, the input data is the image with an amount of random noise added, and the target is the original image. The idea is that the important galaxy information is located in the centre for all images, therefore all background space will be amalgamated and have the noise reduced. Figure 13, below, is an example of the results of the noise reduction.

Unfortunately, we were unable to produce a model this way which removed noise without damaging the characteristic features of galaxies. Figure 14 illustrates an image where the noise reduction removed characteristic galaxy data. As the noise in the images is generally manageable, we decided to use the original images without noise reduction to retain all possible features.

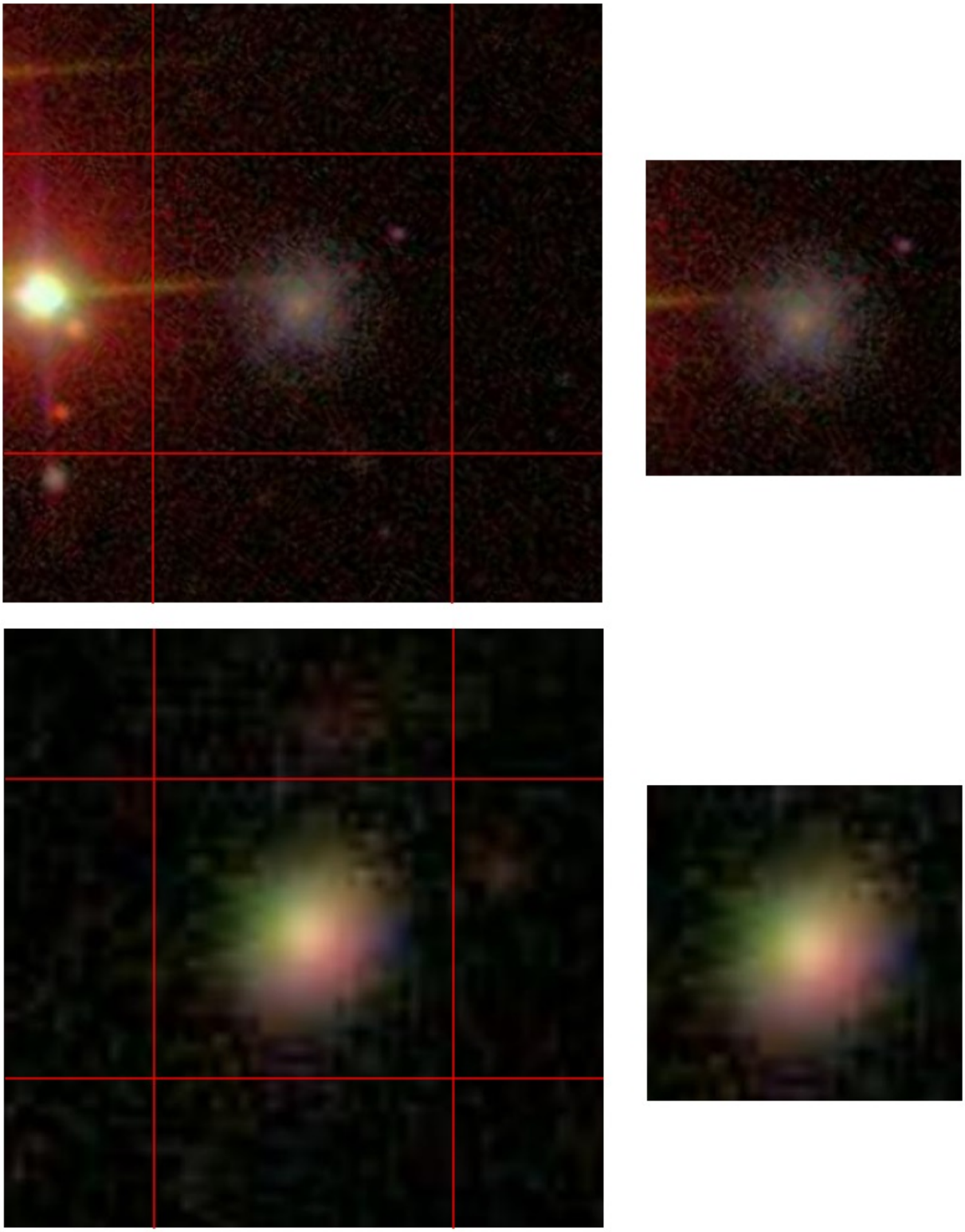

Figure 12: Two examples of galaxy images which were flagged due to potential information loss. It can be seen how image noise and ambient light can result in a flag.

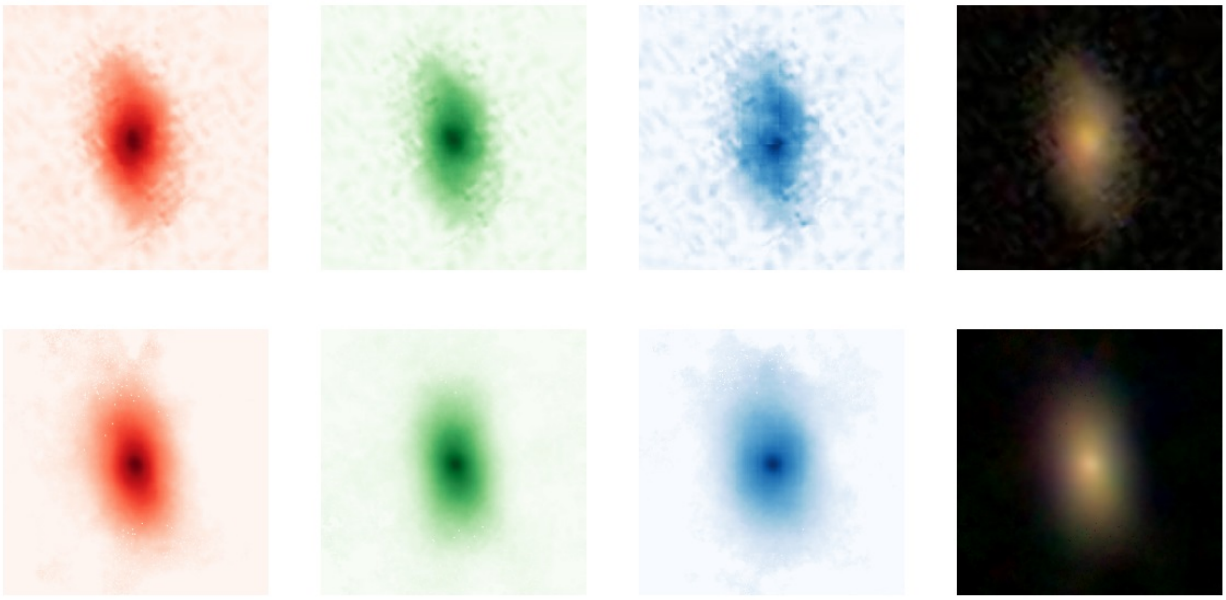

Figure 13: The results of the noise reduction model. The top row is the original image split into the visible red, green, and blue channels, then combined on the far right. The lower row is the result of the noise reduction, again split into the visible red, green, and blue channels.

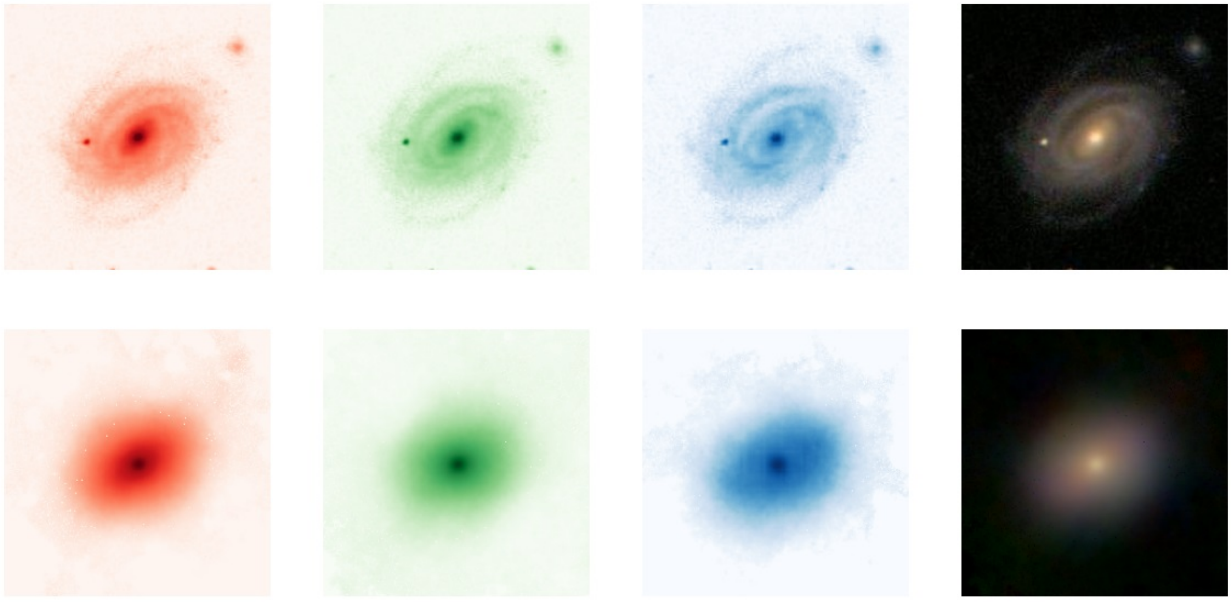

Figure 14: An example of how the noise reduction removes the characteristic tails from a spiral galaxy.

Preliminary Model Creation

The first model was relatively simplistic consisting of three convolutional layers, two fully connected layers and an output layer. After each of the convolutional layer, layers of pooling and dropout were also used. Both of these layers are used to help reduce possible over-fitting and ultimately produce a more reliable model. Each layer utilized a rectified linear unit (ReLU) activation function, with the exception of the output layer, which utilized a sigmoid activation function. The ReLU function "imitates" a linear function and while still allows for the back propagation required for neural network. The sigmoid function was used a our data must fit a two-class classification. The equations for both activation functions can be seen below.

$$\text{ReLU} : h(x) = \max(x, 0)$$

$$\text{Sigmoid} : h(x) = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}}$$

When training the model, as only two classifications were possible, a binary cross-entropy loss function was used. To optimize the training process, the adaptive moment estimation (ADAM) optimizer was used. After training the 10 epochs, the model began to settle on a validation accuracy of 90.9%, with a training accuracy of 96.1%. Comparatively, this is less accurate than the random forest model, with a higher potential for over-fitting, however, it uses no measured parameters of a galaxy which may be favourable in some situations.

VGG-16 Model

For our next model I created a CNN based on the Visual Geometry Group 16 Model (VGG-16). It was originally created at the University of Oxford as part of the 2014 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) competition (Simonyan & Zisserman 2014). This model held the number one spot for a number of years before being beaten out by newer models such as ResNet, Inception, and EfficientNet. Therefore, I figured it would be a good model to test out because it is not to complicated and relatively easy to implement.

The VGG-16 model consists of the following,

1. One block consisting of 2 of Convolution layers (kernel size of 3) with 64 filters and 1 max pooling layer.
2. One block consisting of 2 of Convolution layers (kernel size of 3) with 128 filters and 1 max pooling layer.
3. One block consisting of 2 of Convolution layers (kernel size 3) with 256 filters followed by 1 max pooling layer.
4. Two blocks consisting of 2 Convolution layers with 512 filters (kernel size 3), again followed by 1 max pooling layer.
5. Two fully connected blocks with dropout layers (dropout of 0.5).

6. One binary sigmoidal activation layer.

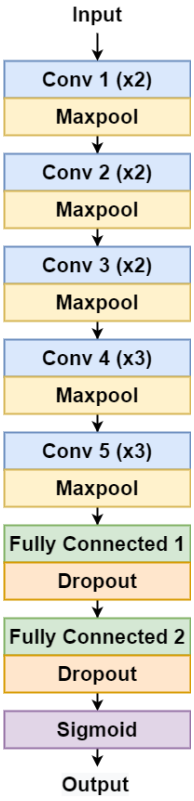

Figure 15: A schematic of the VGG-16 model used in this project. This figure was created in draw.io.

The VGG-16 model is visualized in Figure 15. In total there are 50,378,561 trainable parameters for this model. Additionally, I down-sampled the images to 106×106 pixels as the input instead of the 212×212 images as it did not seem to impact accuracy and made training times faster.

For training we used the 'Adam' optimizer with a learning rate of 1×10^{-6} and the binary cross-entropy loss function. The model was set up to two thirds of the training data divided by the batch size for the training steps and one third of training data divided by the batch size for the number of validation steps per epoch. The model was trained for 25 epochs (batch size 32) with a training time of 250 minutes (4.2 hours), with each epoch taking approximately 300 seconds of run time. The training was done on a local machine with GPU acceleration enabled. The model achieved a training Accuracy of 0.9810 (98.10%) and a validation accuracy of 0.9814 (98.14%) as shown in Figure 16. The training loss was 0.0523 with a validation loss of 0.0508 as shown in Figure 17.

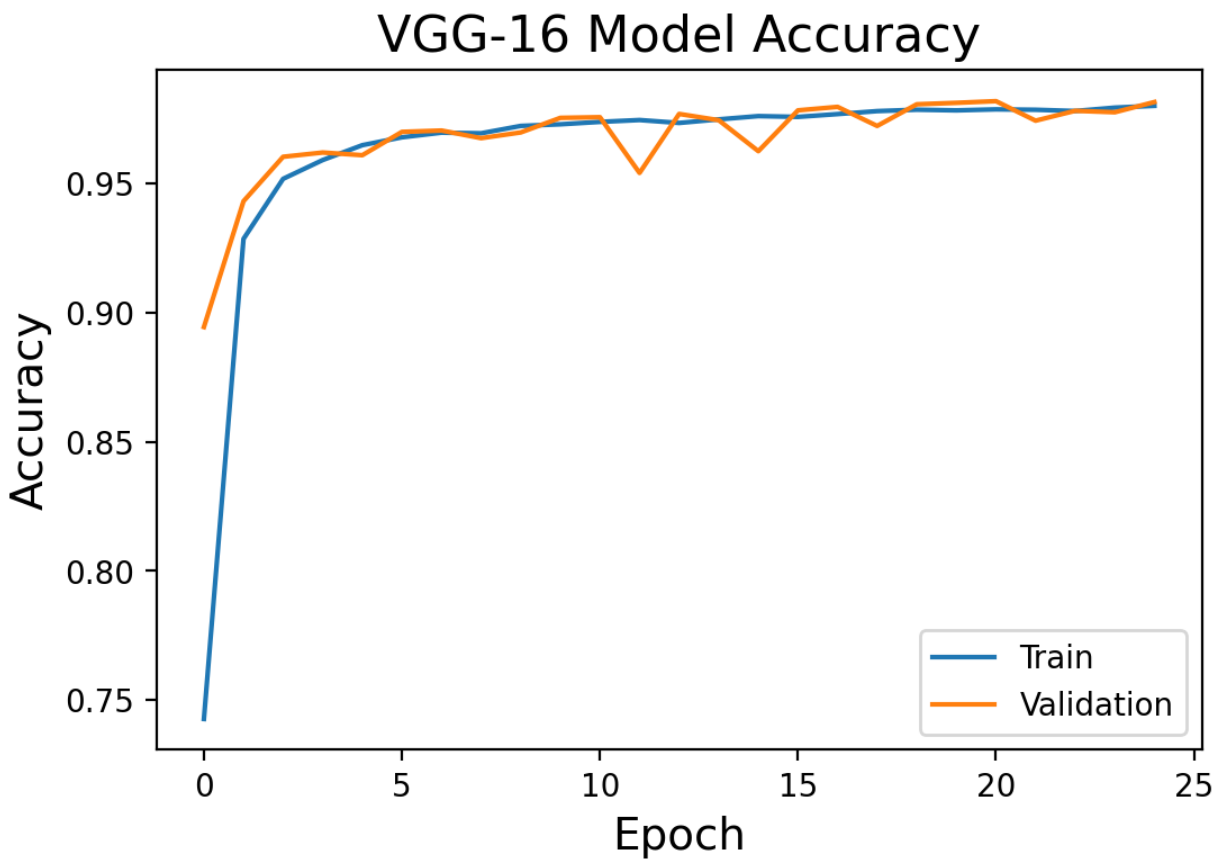

Figure 16: A plot of the training and validation accuracy for the VGG-16 model after 25 epochs. The blue link is the training accuracy and the orange line is the validation accuracy.

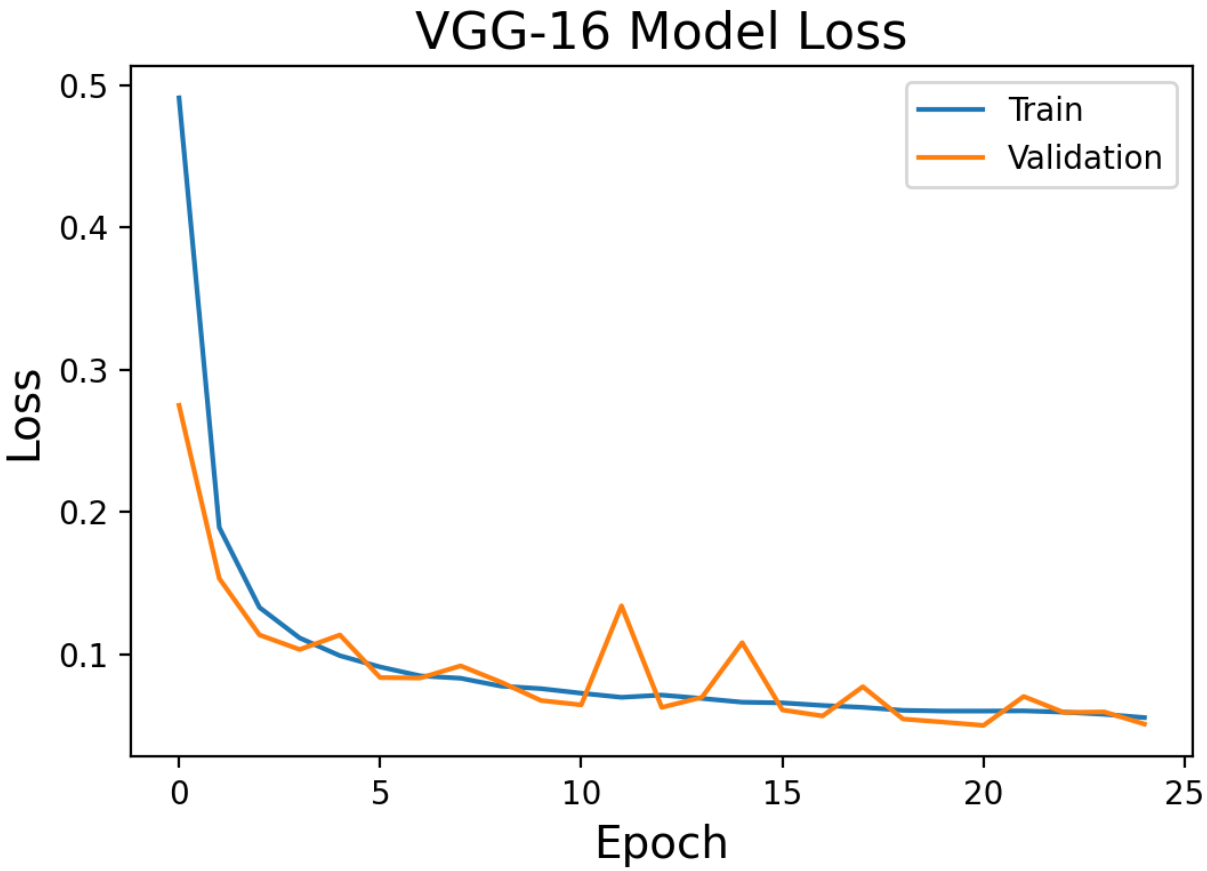

Figure 17: A plot of the training and validation loss for the VGG-16 model.

Predictions

As there is no *correct* classification to hold against the uncertain images, we used our models to predict the classes and compared results between models. Our hope was that the models would classify the images the same way, given the reported accuracy of the models The models we chose to compare were the Random Forest Classifier and the constructed VGG-16 CNN model. Between these two models, they were found to agree on classifications roughly 53% of the time.

Applying the VGG-16 model to the 124,358 unclassable images gave us the following results shown in Figures 18 and 19, as well as in Table 2. The model predicted an additional 52,894 spiral galaxies, compared to the 81,401 confirmed spirals. We also predicted 69,240 elliptical galaxies on top of the already confirmed 33,809 elliptical galaxies in the Galaxy Zoo images. This gives us a combined population of 134,295 spiral galaxies and 103,049 elliptical galaxies as shown in Figure 19.

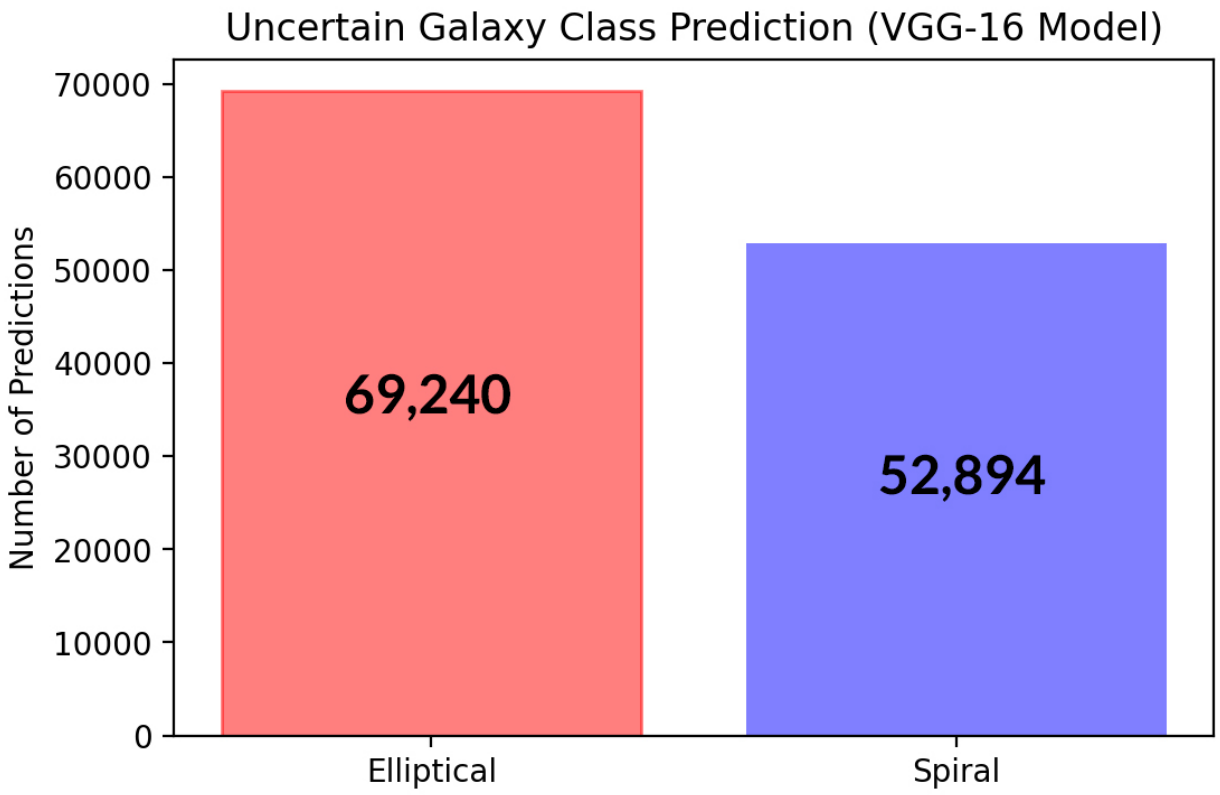

Figure 18: A bar plot of the number of galaxy classes predicted by the VGG-16 model on the images 124,358 marked as 'uncertain'.

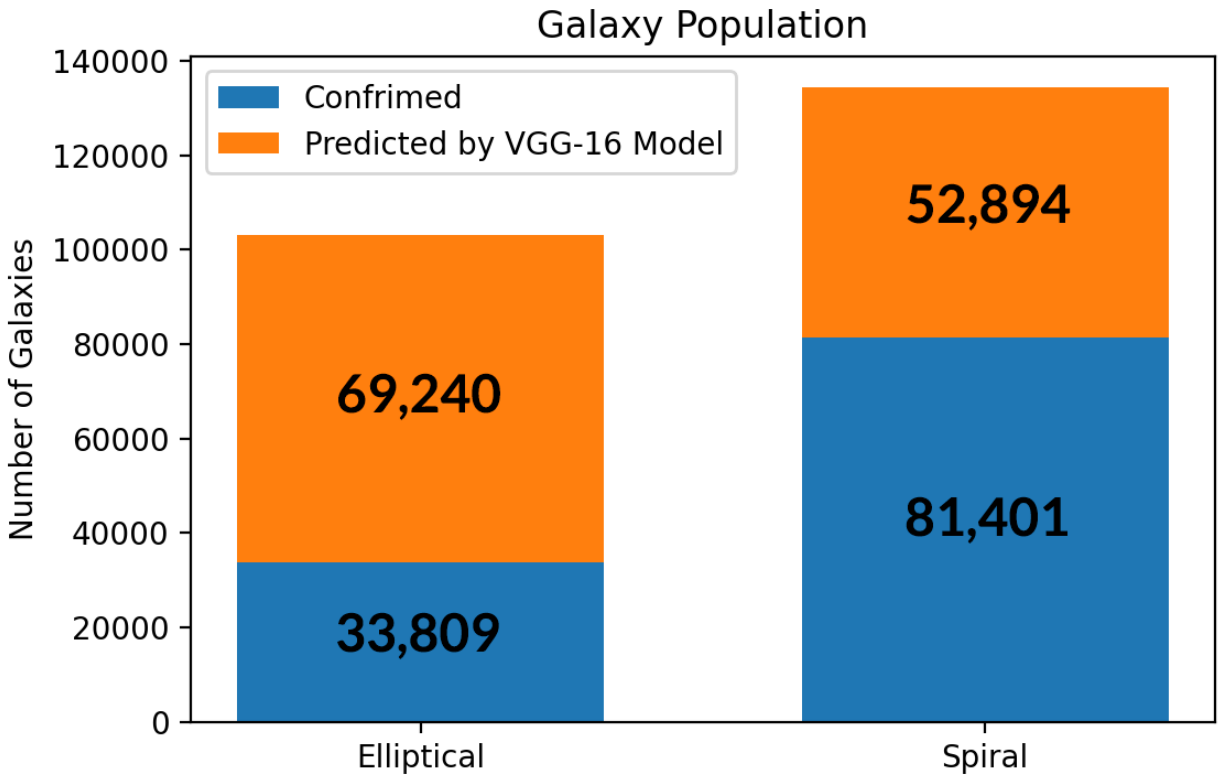

Figure 19: A bar plot showcasing the inferred galaxy population of the Galaxy Zoo images. The blue bars are the confirmed classifications from the catalog and the orange bars are the predicted classifications made by the VGG-16 model.

	Spirals	**Ellipticals**
Confirmed	81,401	33,809
Predicted	52,894	69,240
Total	134,295	103,049

Table 2: Results of the VGG-16 Galaxy Prediction.

Figures 20 and 21 show 4 × 4 mosaics of a random set of the newly classified galaxies to see if they seemed to match up with the predictions. In general, most of the images appear to be visually correct with the prediction but this can sometimes be deceiving for a variety of reasons which will be discussed later. There are a couple of interesting images in the two mosaics. For example, in Figure 20 panel[2] (1,3) was classed as a spiral galaxy but might actually be elliptical. Panels (2,4) and (4,4) are likely an edge on spiral galaxies as the model predicted. In Figure 21 we can see an overlapping pair of galaxies and a giant elliptical galaxy. Panel (4,1) shows an image of galaxy that is classified as spiral but might actually be an elongated elliptical galaxy.

[2]We follow conventional matrix indexing for the mosaic panels.

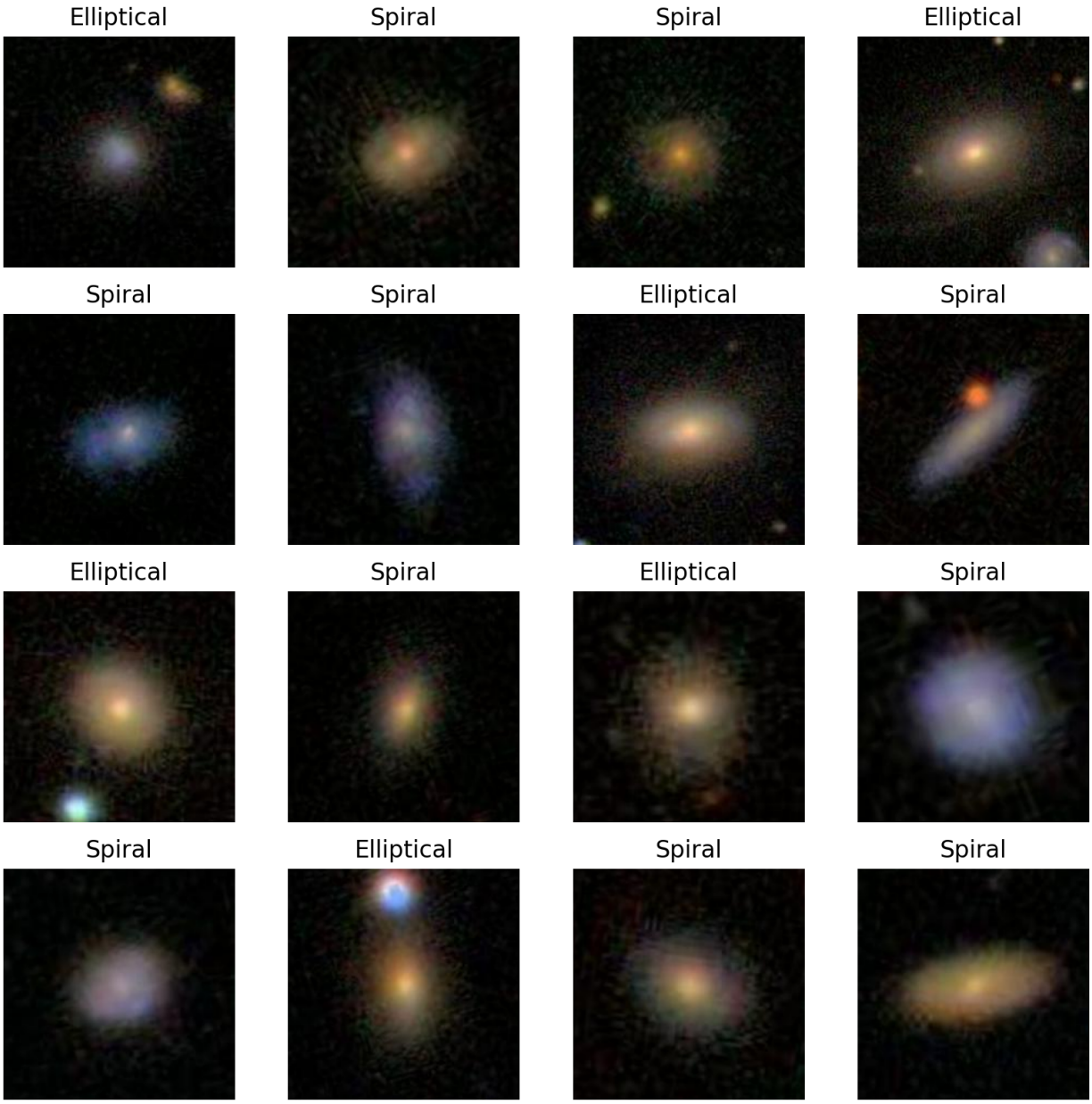

Figure 20: A 4 × 4 mosaic of galaxies and their predicted classification from the VGG-16 model.

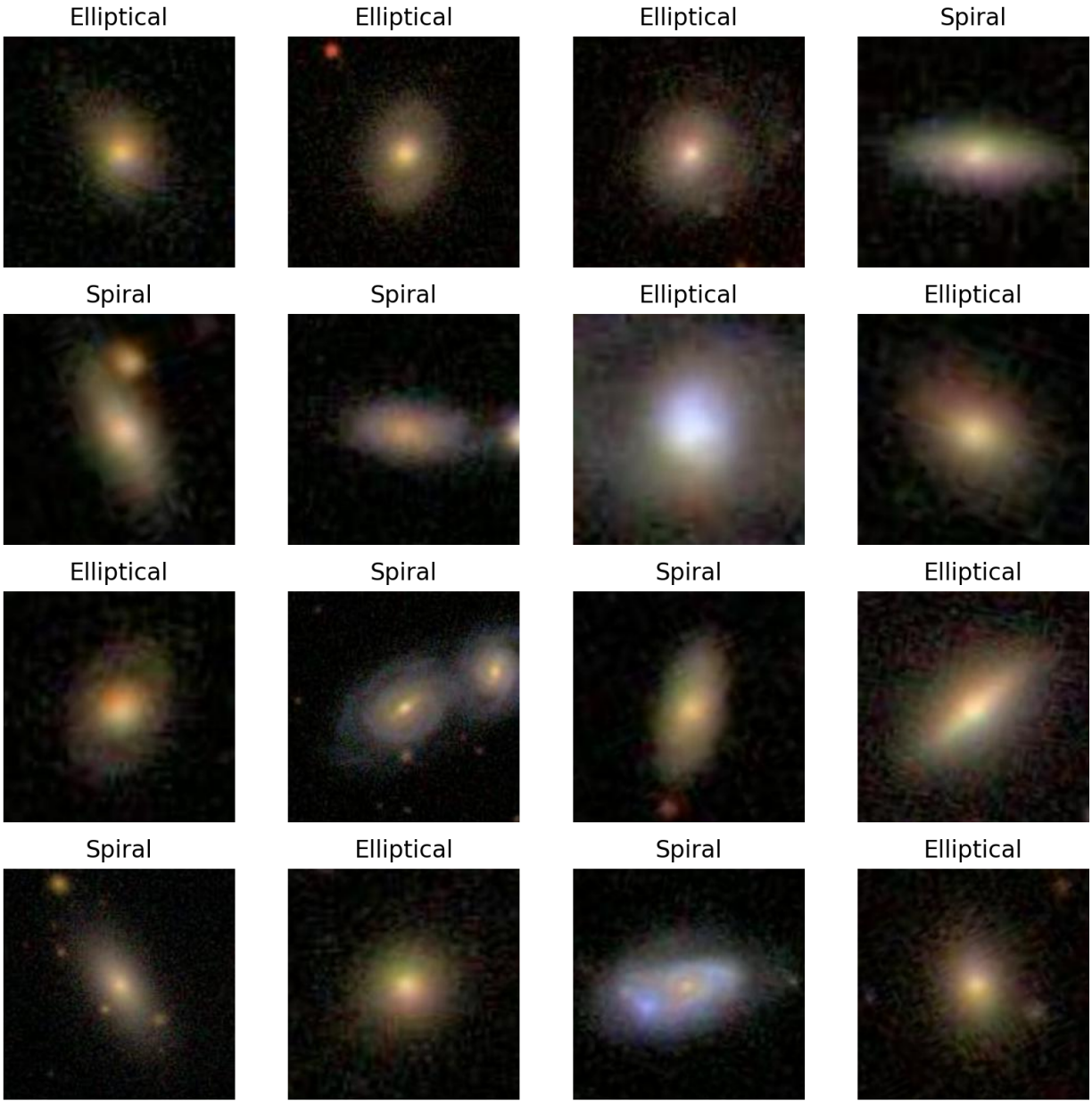

Figure 21: Another 4 × 4 mosaic of galaxies and their predicted classification. There are some interesting objects in panels (2,3), and (3,2). Panel (3,2) contains two overlapping galaxies. There may be another galaxy nearby at the edge of the image in panel (2,2).

As seen in Figure 18, we predict more elliptical galaxies than spiral galaxies. Even though we ultimately predict that there are more spiral galaxies than elliptical galaxies in the whole galaxy Zoo data set (Figure 19) we still find a higher number of elliptical galaxies compared to the literature. It is generally accepted that 70% of galaxies are spiral galaxies (Late-type), and 30% are elliptical galaxies (Early-type). But the model predicts that 55.67% of uncertain are elliptical and 42.53% of the uncertain images contain spiral galaxies. This gives us a total population of 43.42% elliptical and 56.58% Spirals. This is about 13.5% higher than expected, but not unusual,

and may be explained by a few different factors related to the images themselves, the model, and astronomical aspects.

Recall that we identified about 1% of the images with image artifacts or other quality issues affecting classification, such as obscuring starlight or noisy images. Additionally, some images have other objects nearby that may affect classification, such as stars or other galaxies in the foreground or background. We identified about ~ 2,500 of these images in total. Also, the model error means that ~ 2,313 or 1.86% galaxies are misclassified, based on the validation error.

There are also some galaxies that overlap with each other in the image. These aren't merging galaxies but are simply foreground and background galaxies that overlap or otherwise appear close together such as in panel (3,2) of Figure 21. There are a total of 1,990 galaxy pairs in the Galaxy Zoo (Keel et al. 2013). This may have confused the CNN enough to classify the galaxy as the other type. Astrophysically, we know that in clusters galaxy type varies with environment, so perhaps the GZ2 just happened to have more elliptical than the rest of the sky (Goto et al. 2003). It is also possible that other classes ended up being grouped in with the elliptical category, most likely lenticular and irregular galaxies. Lenticular galaxies are thought to occupy about 10% of the galaxy population. Dwarf spirals and irregulars can also roughly appear like an elliptical galaxy sometimes. See the images in Buta (2011) for a review. Overall, it is likely a combination of all of these that contribute to our findings.

Image Size Comparison

To study the effects of changing the pixel counts in our images we created a hand rolled CNN which was kept the same for each image size. This model consisted of the following:

1. One block consisting of a convolutional layer with a filter size of 32 and kernel size of 3. Followed by a 2-D max pooling layer of size (2,2). With a dropout layer of 0.25.
2. The second block consisting of a convolutional layer with a filter size of 64 and kernel size of 3. Followed by a 2-D max pooling layer of size (2,2). With a dropout layer of 0.25.
3. The third block block consisting of a convoluational layer with a filter size of 128 and a kernel size of 3. Followed by a 2-D max pooling layer of size (2,2). With a dropout layer of 0.25.
4. The fourth block which has a flattening layer and a dense layer of size 512. With a dropout layer of 0.5.
5. The fifth and final layer which is a dense layer of 1, using sigmoid activation.

All models were trained using the Adams optimizer with the default learning rate over 15 epochs. The image sizes that were used for this analysis were 212x212, 156x156, 106x106, 53x53, and 26x26. First we ran the full sized 212x212 images through the hand rolled CNN. A sample of the images used can be seen below in Figure 22.

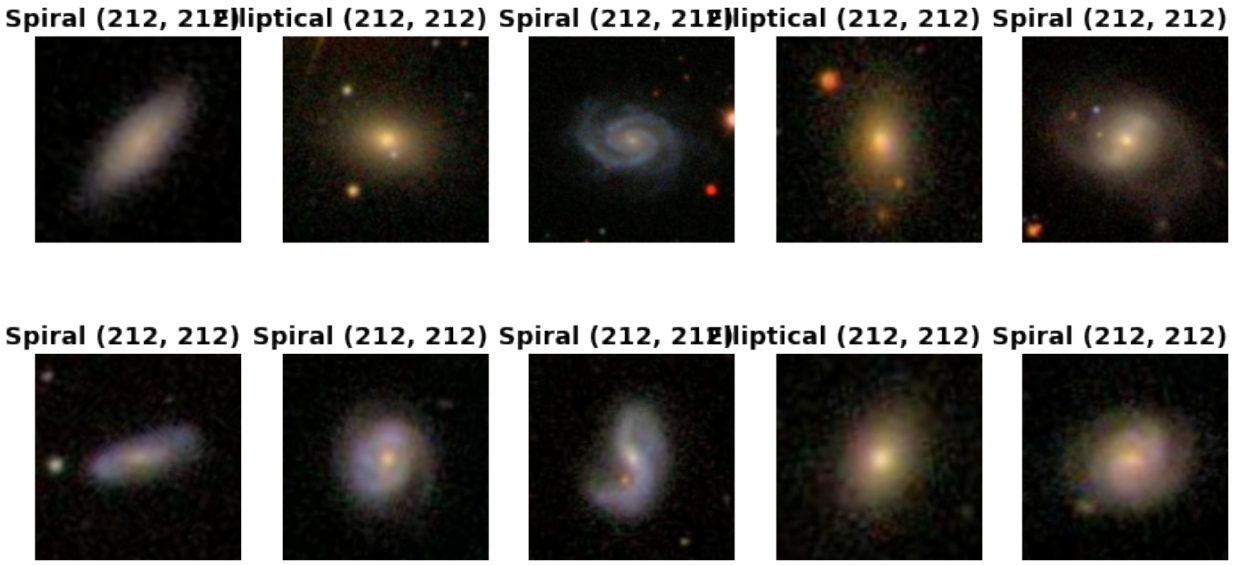

Figure 22: A sample of 212x212 images used.

This produced a model that had 37,843,009 trainable parameters. Using Tensorflow GPU we had a run time of 6554 seconds or 1.82 hours. We obtained a training accuracy of 0.9616 with a training loss of 0.1032 computed using the binary cross entropy loss function. The validation accuracy was 0.9521 with a validation loss of 0.1228. We felt as though this was a good result even with the substantial training time.

Below in Figure 23 is a sample of the 156x156 images that were used for our analysis.

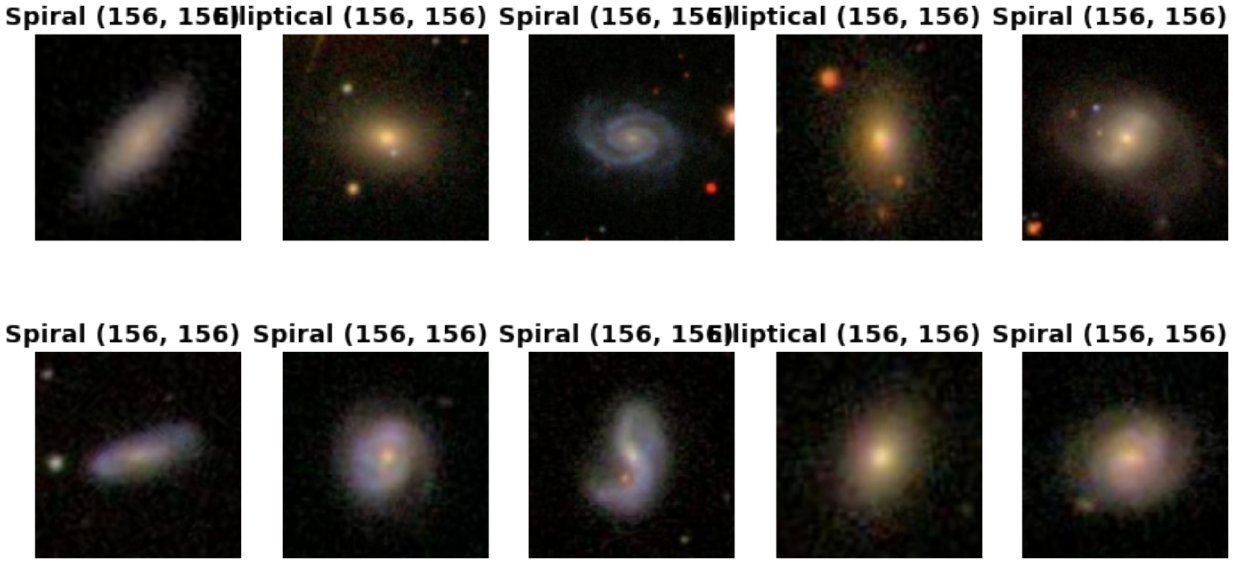

Figure 23: Sample of 156x156 images used.

This produced a model that had 19,034,177 trainable parameters. The total run time for the

15 epochs was 4059 seconds or 1.13 hours with a computed training accuracy of 0.9603 and training loss of 0.1071. The validation accuracy was 0.9632 with a validation loss of 0.1097. This was a slightly better result by all metrics compared to the 212x212 as we saw an improvement of 1% in our validation accuracy along with a 1% decrease in the validation loss. Couple these improvements with the substantially lower run time and the 156x156 images appeared to be far more preferable than using our full sized images.

The next model used the 106x106 images. A sample of these can be seen below in Figure 24.

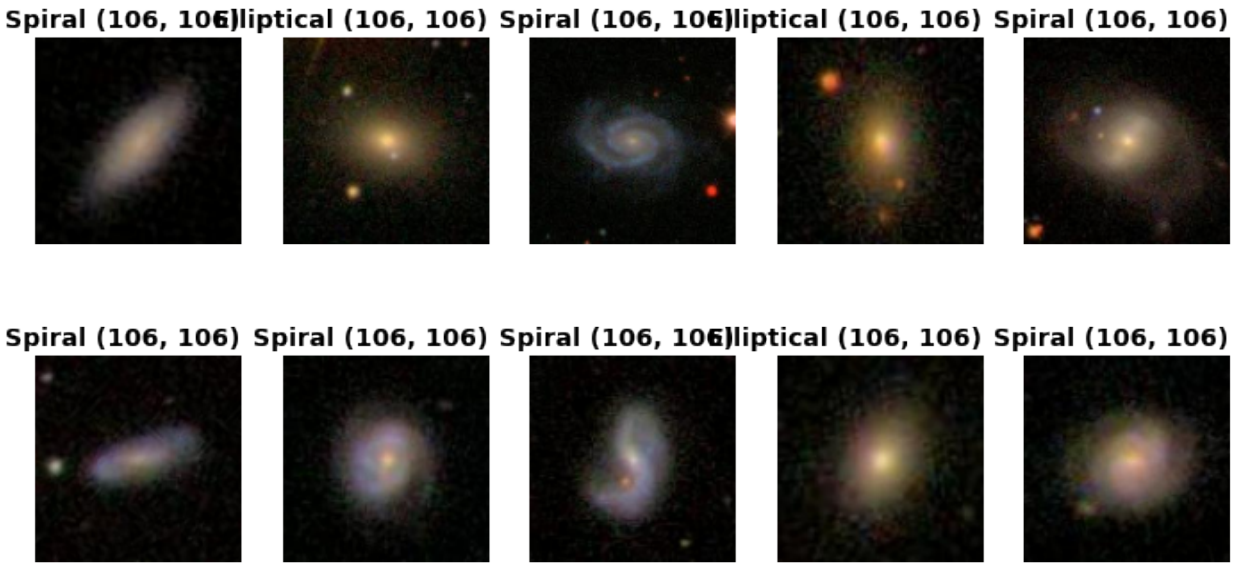

Figure 24: Sample of the 106x106 images used.

This produced a model with 8,024,129 trainable parameters. The total run time for the 15 epochs was 2422 seconds or 40.3 minutes. We computed a training accuracy of 0.9696 with a training loss of 0.0807. The computed validation accuracy was 0.9706 with a validation loss of 0.0778. Again we see a reasonable improvement compared to the previous image sizes.

The fourth model trained used images of size 53x53 and a sample of them can be seen in Figure 25

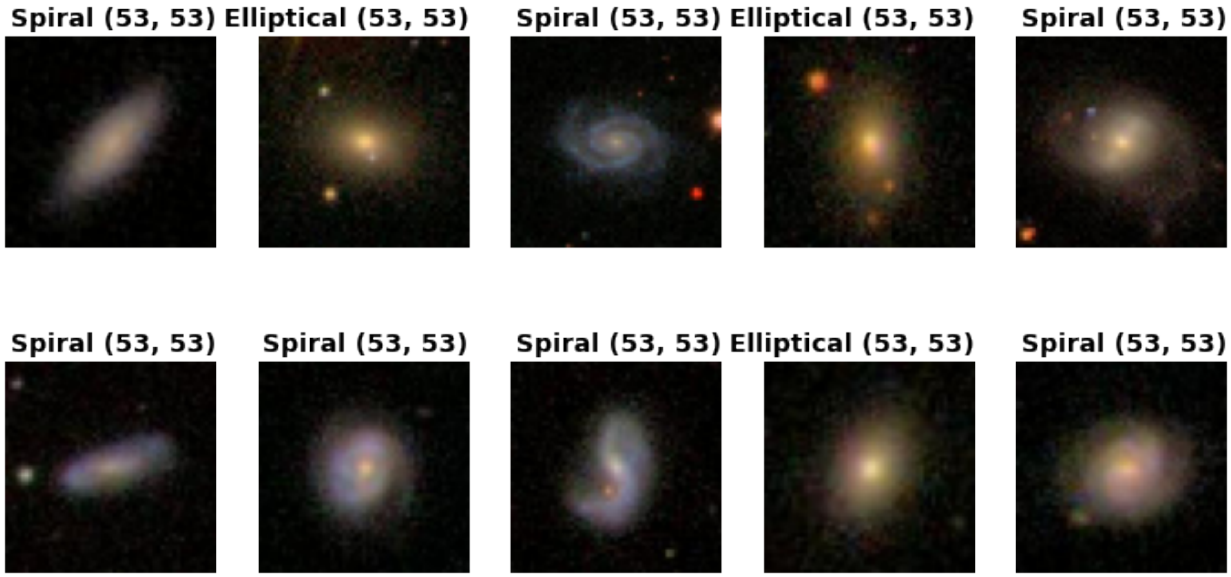

Figure 25: Sample of 53x53 images used.

Using the 53x53 sized images we produced a model that contained 1,142,849 trainable parameters. The total run time for the 15 epochs was 998 seconds or 16.6 minutes. We computed a training accuracy of 0.9736 with a training loss of 0.0749. With a validation accuracy of 0.9756 and validation loss of 0.0652. This model is again an improvement on the previous iteration finishing with a much faster run time for the same amount of epochs with better validation accuracy and loss.

The final model we produced for this analysis was using the 26x26 sized images in Figure 26

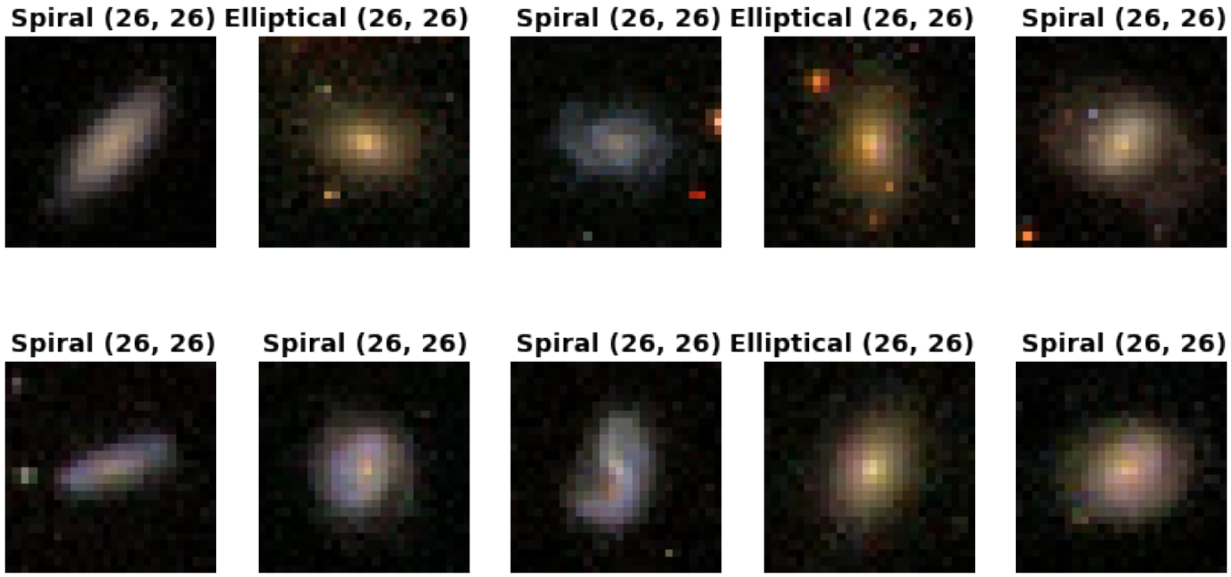

Figure 26: Sample of 26x26 images used.

With the 26x26 sized images we produced a model with 159,809 trainable parameters. This model had a total run time of 685 seconds or 11.4 minutes for the 15 epochs. We computed a training accuracy of 0.9650 with a training loss of 0.0938. The validation accuracy was 0.9651 and the validation loss was 0.0940. The accuracy's and loss values are all slightly worse than those for the 53x53 images. This could suggest that we have finally begun to lose pixel data with such a small number of pixels present in the images. It is also worth noting that decreasing the trainable parameters by 10x only gave us a performance increase of 30% as we start to see substantial diminishing returns in run times with these lower pixel counts.

Conclusions

The multinomial Naive Bayes model produced the lowest model accuracy of all tested models. Its limitation is likely that it cannot handle negative values. Comparatively, the random forest model produced a promising testing accuracy, and once tuned to reduce or eliminate over-fitting, becomes more favourable. While it is possible to feed image data to these model, we decided to only include the numerical measurements collected within the listed databases.

While there is worth to the random forest model, the work required to collect, measure, and calculate all the required parameters likely outweigh the resources required to collect a single image of the galaxy. The measurements are usually done long after the image is created in the processing pipeline.

Using the images in the convolutional neural networks first required some data conditioning and management. Attempts were made to reduce the noise within the images but no constructed models appeared to produce the desired results.

With the images, several convolutional neural networks were produced. The first, and most simplistic, produced reasonable accuracy's but was outperformed by all models except the Naive Bayes. The more complicated VGG-16 model had the best performance out of all the models in this project, achieving a validation accuracy of ~ 98%. However, when predicting on the images marked as uncertain the VGG-16 model only agreed with the random forest classifier ~ 53% of the time. This suggests that one of the models is faulty in some way, or that the data is difficult to interpret. But due to the many variables and other unknowns that a play a role in astronomical processes, it is hard to pinpoint the exact reason or reasons as to why the random forest and CNN models did not agree with each other. It is concluded that likely a combination of model design and astronomical related mechanisms played a role in this discrepancy.

The image size that produced the best results for our hand rolled CNN were the 53x53 sized images which decreased our total run time from 1.8 hours to 16.6 minutes and provided the best validation accuracy and loss out of all the image sizes that we used for our analysis. We suspect that the reason for this is that the images are already at such a low resolution that single pixels did not contain much information that was different than neighbouring pixels. Thus when we decreased the pixel count we were not losing any valuable information from the images. This could have lead to less over-fitting in the models to useless information and helped improve our accuracy with smaller image sizes. It also allowed us to shorten the run time by a substantial amount.

Future Work

Machine learning in general is pretty new to astronomy, only becoming popular around 2014 (Ivezic 2014). Now, traditional machine learning techniques are now being used in various data processing pipelines and analysis. An example of this is classifying stars, galaxies or quasars. Deep learning is still more or less in its infancy, we are still experimenting with different applications of them. This includes, identifying anomalies in images, various classification problems similar to what we are doing, and detecting and predicting photometric redshifts of objects (Pasquet-Itam & Pasquet 2018). One of the goals of the Galaxy Zoo project was to get enough classification data to use to experiment and train on deep learning techniques, as well as detect other patterns that could be used in classification.

In the future, we would like to train a random forest classifier on the Galaxy Zoo 2 Classification System rather than on the simpler binary Galaxy Zoo 1 classification that we used here. The Galaxy Zoo 2 has 32 galaxy classes that better represent those shown in Figure 1 or 2. We would also like to train a CNN model to classify the images based on the classifications provide in the GZ2 catalogue.

References

Bernstein, G. M. & Jarvis, M. (2002), 'Shapes and Shears, Stars and Smears: Optimal Measurements for Weak Lensing', *Astronomical Journal* **123**(2), 583–618.

Blanton, M. R. et al. (2017), 'Sloan Digital Sky Survey IV: Mapping the Milky Way, Nearby Galaxies, and the Distant Universe', *Astronomical Journal* **154**(1), 28.

Buta, R. J. (2011), 'Galaxy Morphology', *arXiv e-prints* p. arXiv:1102.0550.

Buta, R. J. et al. (2015), 'A Classical Morphological Analysis of Galaxies in the Spitzer Survey of Stellar Structure in Galaxies (S4G)', *Astrophysical Journal Supplement* **217**(2), 32.

Conselice, C. J. et al. (2016), 'The Evolution of Galaxy Number Density at z 8 and its Implications', *The Astrophysical Journal* **830**(2), 83.

Curtis, H. D. (1917), 'Novae in the spiral nebulae and the island universe theory', *Publications of the Astronomical Society of the Pacific* **29**, 206.

Darg, D. W. et al. (2010), 'Galaxy Zoo: the fraction of merging galaxies in the SDSS and their morphologies', *Monthly Notices of the Royal Astronomical Society* **401**(2), 1043–1056.

de Vaucouleurs, G. (1948), 'Recherches sur les Nebuleuses Extragalactiques', *Annales d'Astrophysique* **11**, 247.

de Vaucouleurs, G. (1959), 'Classification and Morphology of External Galaxies.', *Handbuch der Physik* **53**, 275.

Dyson, F. W., Eddington, A. S. & Davidson, C. (1920), 'A Determination of the Deflection of Light by the Sun's Gravitational Field, from Observations Made at the Total Eclipse of May 29, 1919', *Philosophical Transactions of the Royal Society of London Series A* **220**, 291–333.

Einstein, A. (1905), 'On the electrodynamics of moving bodies', *Annalen Phys.* **17**, 891–921.

Eisenstein, D. J. et al. (2011), 'SDSS-III: Massive Spectroscopic Surveys of the Distant Universe, the Milky Way, and Extra-Solar Planetary Systems', *Astronomical Journal* **142**(3), 72.

Goto, T. et al. (2003), 'The morphology-density relation in the Sloan Digital Sky Survey', *Monthly Notices of the Royal Astronomical Society* **346**(2), 601–614.

Gunn, J. E. et al. (2006), 'The 2.5 m Telescope of the Sloan Digital Sky Survey', *Astronomical Journal* **131**(4), 2332–2359.

Hart, R. E. et al. (2016), 'Galaxy Zoo: comparing the demographics of spiral arm number and a new method for correcting redshift bias', *Monthly Notices of the Royal Astronomical Society* **461**(4), 3663–3682.

Hubble, E. (1926), 'No. 324. Extra-galactic nebulae.', *Contributions from the Mount Wilson Observatory / Carnegie Institution of Washington* **324**, 1–49.

Hubble, E. (1929), 'A relation between distance and radial velocity among extra-galactic nebulae', *Proceedings of the National Academy of Sciences* **15**(3), 168–173.

Ivezic, Z. (2014), *Statistics, data mining, and machine learning in astronomy : a practical Python guide for the analysis of survey data*, Princeton University Press, Princeton, N.J.

Keel, W. C. et al. (2013), 'Galaxy Zoo: A Catalog of Overlapping Galaxy Pairs for Dust Studies', *Publications of the Astronomical Society of the Pacific* **125**(923), 2.

Lintott, C. J. et al. (2008), 'Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey', *Monthly Notices of the Royal Astronomical Society* **389**(3), 1179–1189.

Lintott, C. et al. (2011), 'Galaxy Zoo 1: data release of morphological classifications for nearly 900 000 galaxies', *Monthly Notices of the Royal Astronomical Society* **410**(1), 166–178.

Morgan, W. W. & Osterbrock, D. E. (1969), 'On the Classification of the Forms and the Stellar Content of Galaxies', *Astronomical Journal* **74**, 515.

Pasquet-Itam, J. & Pasquet, J. (2018), 'Deep learning approach for classifying, detecting and predicting photometric redshifts of quasars in the Sloan Digital Sky Survey stripe 82', *Astronomy and Astrophysics* **611**, A97.

Schawinski, K. et al. (2010), 'Galaxy Zoo: The Fundamentally Different Co-Evolution of Supermassive Black Holes and Their Early- and Late-Type Host Galaxies', *Astrophysical Journal* **711**(1), 284–302.

Shapley, H. & Curtis, H. D. (1921), 'The Scale of the Universe', *Bulletin of the National Research Council, Vol. 2, Part 3, No. 11, p. 171-217* **2**, 171–217.

Simonyan, K. & Zisserman, A. (2014), 'Very Deep Convolutional Networks for Large-Scale Image Recognition', *arXiv e-prints* p. arXiv:1409.1556.

Sparke, L. (2007), *Galaxies in the universe : an introduction*, Cambridge University Press, Cambridge New York.

Willett, K. W. et al. (2013), 'Galaxy Zoo 2: detailed morphological classifications for 304 122 galaxies from the Sloan Digital Sky Survey', *Monthly Notices of the Royal Astronomical Society* **435**(4), 2835–2860.

York, D. G. et al. (2000), 'The Sloan Digital Sky Survey: Technical Summary', *Astronomical Journal* **120**(3), 1579–1587.