Georgia Institute of Technology
CX 4240 Computing For Data Analysis

Galaxy Morphology Classification Using CNN

Karen Loscocco and Yunqing Jia

July 2019

Contents

1 Introduction

1.1 Motivation

The universe is approximately 13.8 billion years old, while the Earth has only been around for 4.54 billion years. As the universe continues to expand, questions remain about how galaxies formed and evolved. To solve this puzzle of the cosmic structure, understanding the distribution and types of galaxies as a function of their shapes and sizes is crucial. As the technology, specifically telescopes, for capturing such phenomena and returning images to scientists on Earth, improves, the dataset increases exponentially. Traditionally, galaxy identifications were achieved by crowdsourcing. However, as the datasets grow to the scale of hundreds of millions of galaxies, manual identification becomes less feasible. This project aims to analyze photos of galaxies and to train an algorithm that generates automated metrics and classifies them into two different classes, spiral and elliptical.

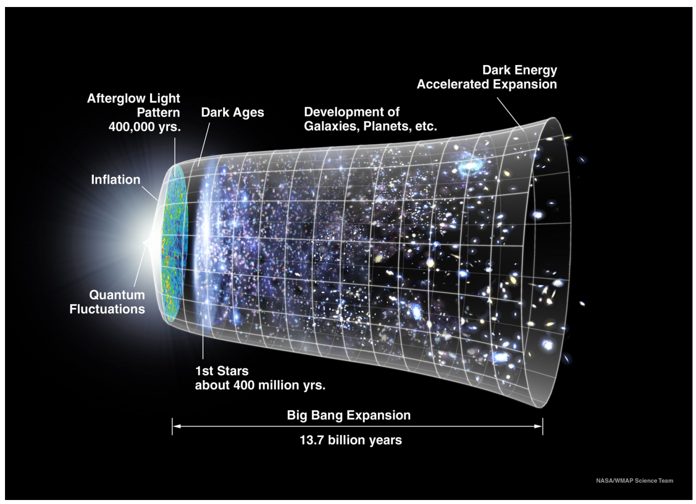

Figure 1.1: Timeline of the Cosmos

2 Problem Definition & Project Scope

The problem we are trying to solve is mapping an already classified set of JPEG images into two different categories. We decided to implement our solution using a Convolutional Neural Network (CNN) since is is widely used when dealing with image data. However, this method does have its drawbacks, the most notorious one being the problem of overfitting, which is inevitable with the large number of parameters generated in the neural network. Dimensionality reduction methods such as principal component analysis (PCA) can be used to counteract such effects, and other machine learning methods such as ridge regression and random forest can be used to generate a baseline for us to evaluate our model's performance. The data preprocessing and the different approaches are documented in detail in the Methodology section.

3 Methodology

3.1 Image Preprocessing

The dataset, acquired from this Kaggle competition, includes 61,578 JPG images that are 424x424 pixels in size, as well as their probability distribution assignments based on 37 crowdsourced classifications. By examining the dataset, most images had a large amount of black margins around the area of interest. Because of this, we decided to crop all images to 100x100 pixels and converted them to grayscale to reduce the computational cost.

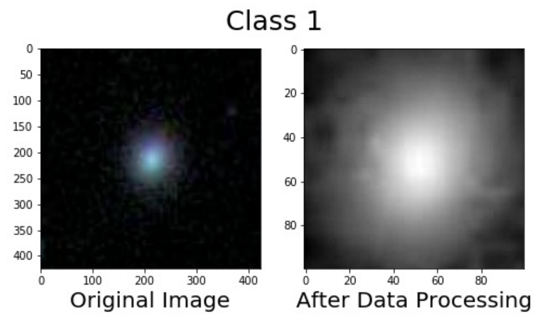

Figure 3.1: Original to Cropped, Grayscale Images Class 1

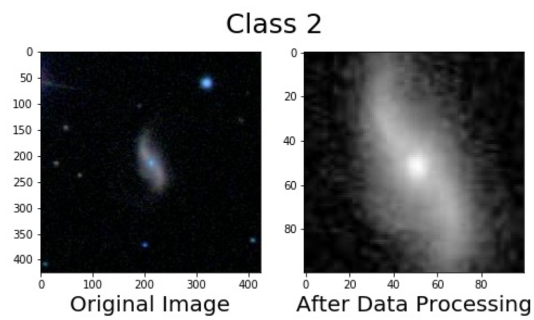

Figure 3.2: Original to Cropped, Grayscale Images Class 2

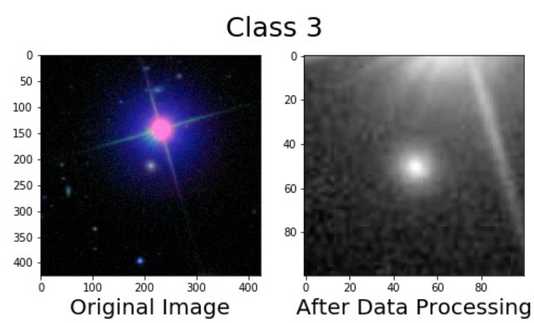

Figure 3.3: Original to Cropped, Grayscale Images Class 3

The original 37 class probability assignments are based on a decision tree, which is comprised of 11 classification tasks, shown in Figure (3.4). All classes but those from the first task of the decision tree are composite probabilities which emphasize the high-level, large-scale morphology categories.

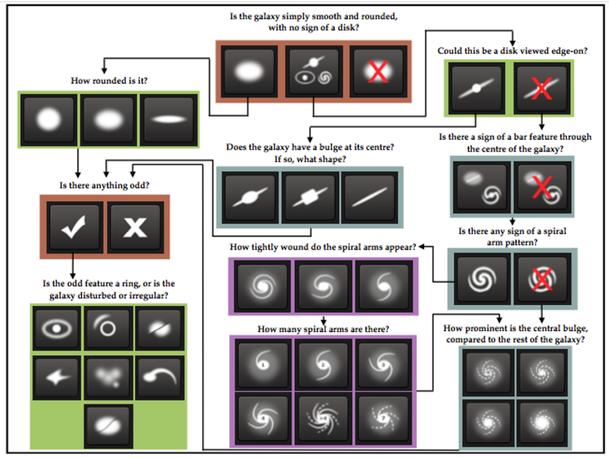

Figure 1. Flowchart of the classification tasks for GZ2, beginning at the top centre. Tasks are colour-coded by their relative depths in the decision tree. Tasks outlined in brown are asked of every galaxy. Tasks outlined in green, blue, and purple are (respectively) one, two or three steps below branching points in the decision tree. Table 2 describes the responses that correspond to the icons in this diagram.

Task	Question	Responses	Next
01	*Is the galaxy simply smooth and rounded, with no sign of a disk?*	smooth	07
		features or disk	02
		star or artifact	**end**
02	*Could this be a disk viewed edge-on?*	yes	09
		no	03
03	*Is there a sign of a bar feature through the centre of the galaxy?*	yes	04
		no	04
04	*Is there any sign of a spiral arm pattern?*	yes	10
		no	05
05	*How prominent is the central bulge, compared with the rest of the galaxy?*	no bulge	06
		just noticeable	06
		obvious	06
		dominant	06
06	*Is there anything odd?*	yes	08
		no	**end**
07	*How rounded is it?*	completely round	06
		in between	06
		cigar-shaped	06
08	*Is the odd feature a ring, or is the galaxy disturbed or irregular?*	ring	**end**
		lens or arc	**end**
		disturbed	**end**
		irregular	**end**
		other	**end**
		merger	**end**
		dust lane	**end**
09	*Does the galaxy have a bulge at its centre? If so, what shape?*	rounded	06
		boxy	06
		no bulge	06
10	*How tightly wound do the spiral arms appear?*	tight	11
		medium	11
		loose	11
11	*How many spiral arms are there?*	1	05
		2	05
		3	05
		4	05
		more than four	05
		can't tell	05

Table 2. The GZ2 decision tree, comprising 11 tasks and 37 responses. The 'Task' number is an abbreviation only and does *not* necessarily represent the order of the task within the decision tree. The text in 'Question' and 'Responses' are displayed to volunteers during classification, along with the icons in Figure 1. 'Next' gives the subsequent task for the chosen response.

Figure 3.4: Decision Tree Decision Tasks

We encountered difficulties enforcing composite constraints on the output layer of our algorithms, thus, we modified our objective by focusing on only the first task of the decision tree, which is composed of 3 classes, spiral, elliptical, and merging galaxies. The results from our algorithms build a foundation for future expansion of the objective.

We then analyzed the distribution of probability assignments for the remaining 3 classes, and discovered that a large portion of data did not have a high confidence in one single category. We filtered out data points with a confidence below 0.8 and plotted a histogram, seen in Figure (3.5). This procedure reduced our data set from 61,578 to 24,273 images.

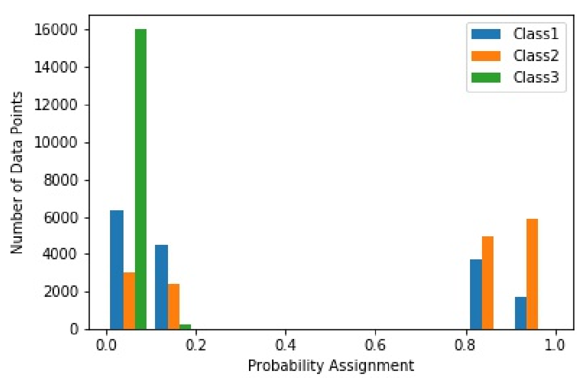

Figure 3.5: Histogram of Probability Distributions for 3 Classes

We also discovered that there are very few samples with a highly confident assignment to the third class. The third class corresponded to merging galaxies, which are rare in this particular dataset. This led us to reduce our objective to a binary classification problem.

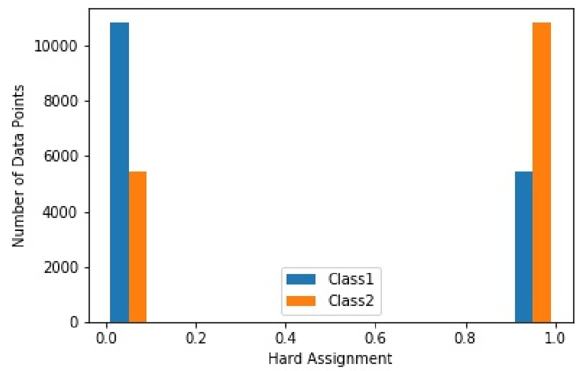

Figure 3.6: Histogram of Hard Assignment Distributions for 2 Classes

3.2 Dimensionality Reduction

We tested several machine learning algorithms with the preprocessed images, but the results were only marginally better than the accuracy of random guessing. The computational cost was also very high, which made it difficult for us to experiment with different models and tweak parameters. To combat this challenge, we reduced the dimension of our dataset using Principal Component Analysis (PCA). The processed image originally had 10,000 components (100x100 pixels). After performing PCA, each data entry was reduced to only 17 components, which covered 95% of the variance. We then used the feature weights generated from PCA to reconstruct images to the cropped size so we can pass them back into our ML models. Figure (3.7) shows images before and after PCA, and Figure (3.8) shows the principal components projected into three dimensions.

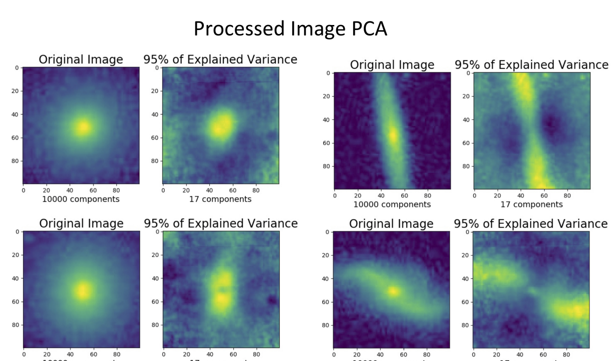

Figure 3.7: Preprocessed to PCA Reconstructed Image

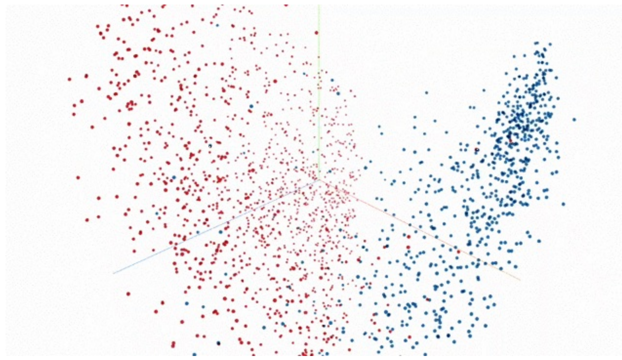

Figure 3.8: Tensorboard Projector - PCA

We also performed PCA on the original 424x424x3 images to validate that the preprocessed images preserved the most important features that best explained the variances in the data. An example of the resulting image is shown in Figure (3.9).

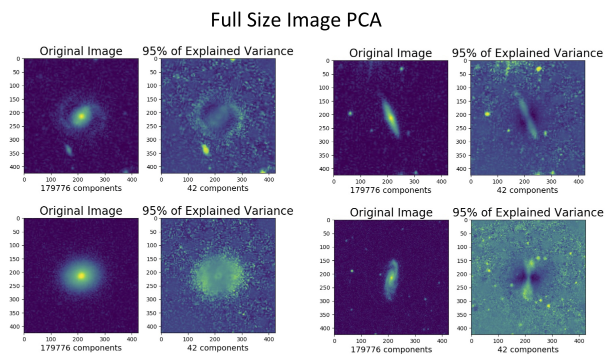

Figure 3.9: Full Size Image to PCA Reconstructed Image

To help visualize the higher dimensional data, we include the following snapshots of the Tensorboard Projector t-distributed Stochastic Neighbor Embedding process.

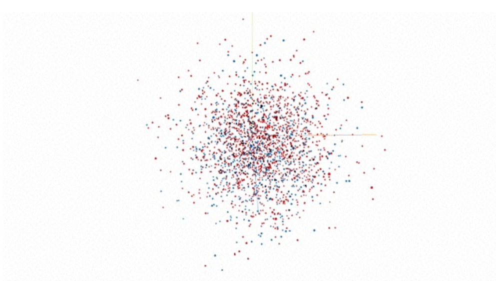

Figure 3.10

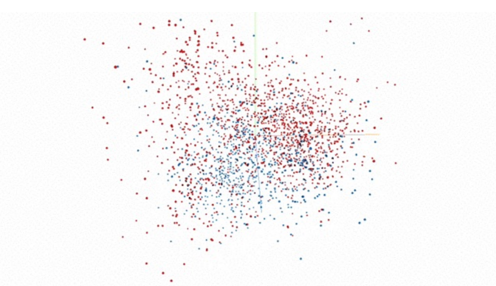

Figure 3.11

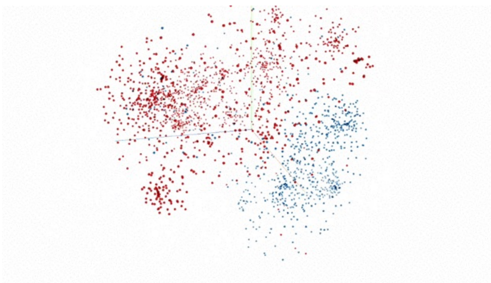

Figure 3.12

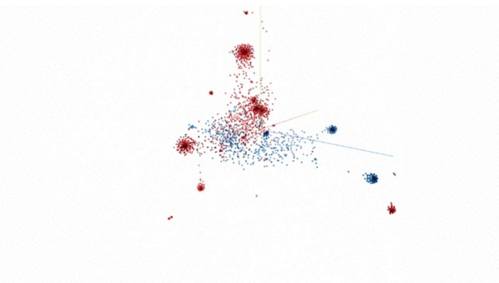

Figure 3.13

4 Sprinkle in Some Machine Learning

4.1 Benchmarks

To set a baseline for the performance of our CNN, we fitted the data using Logistic Regression, Ridge Regression, and Random Forest. The dataset was split into 21,845 training points and 2,428 testing points. For the logistics regression, we used 'liblinear' as our solver, and for the ridge regression, we set alpha to 100. The confusion matrices and the accuracy scores for all three models are shown in the following figure.

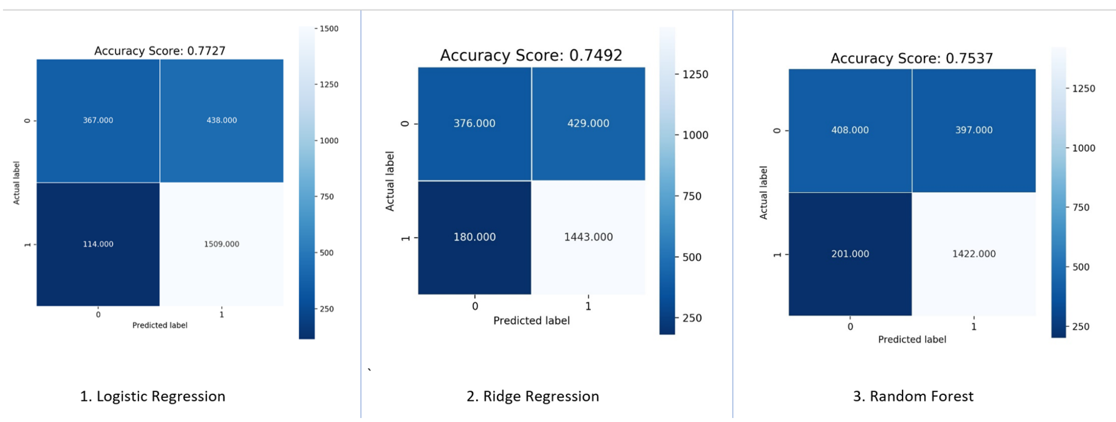

Figure 4.1: Confusion Matrices for the Baseline Models

4.2 Convolutional Neural Network (CNN)

The best model we found is a variation of a model that was used to train the MNIST dataset. We tweaked the sequence of layers and their parameters, the architecture of our final model is in Figure (4.2).

The input is presented to the model in the form of grayscale 100x100 image parts after dimensionality reduction (PCA).

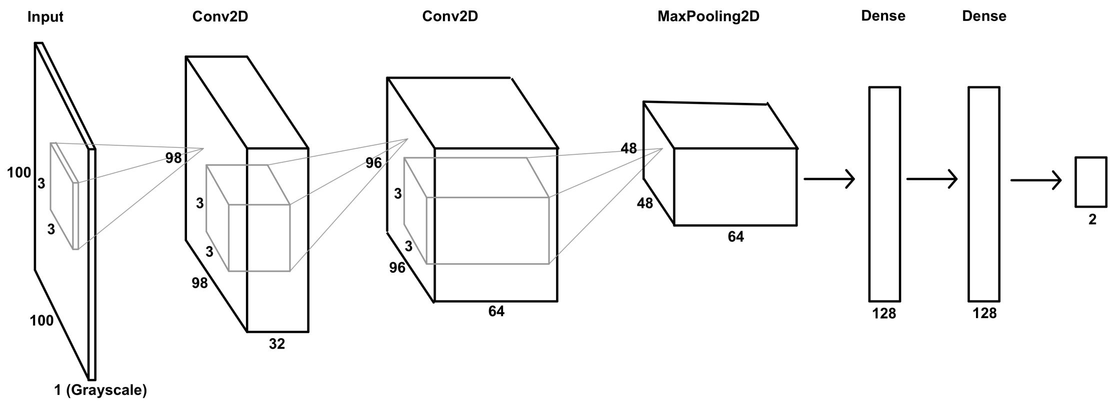

Figure 4.2: CNN Model Architecture

The model had five layers including two convolutional layers and two dense layers. All convolutional layers included a ReLU activation to capture the nonlinearity (i.e. f(x) = max(x, 0)). The second convolutional layer is followed by 2x2 max-pooling. The number of nodes and the kernel size for each layer is labeled in Figure (4.2). We trained the CNN with ADADELTA as the optimizer, which is an adaptive learning rate method for gradient descent. We defined our loss function to be categorical cross-entropy, which works well with hard assignments. And for our evaluation metrics, we computed the mean absolute error and the accuracy value on our training and validation data.

5 Results

The model achieved high accuracy: 98.77% for training and 95.74% validation.

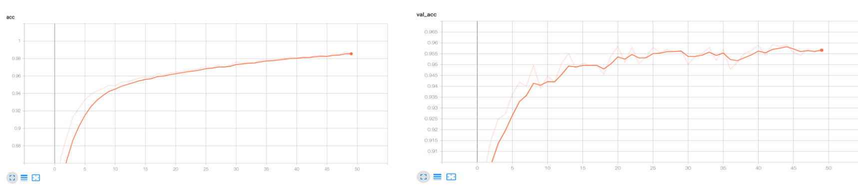

Figure 5.1: Training and Validation Accuracy over 50 Epochs

The mean absolute error and the accuracy value converged after 50 epochs.

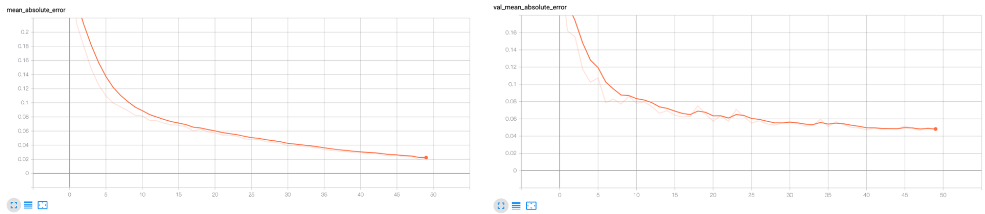

Figure 5.2: Training and Validation Mean Absolute Error over 50 Epochs

The validation loss started to increase again after about 15 epochs, which could be a sign of overfitting.

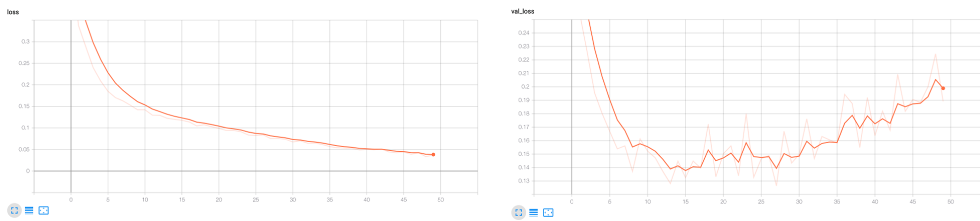

Figure 5.3: Training and Validation Loss over 50 Epochs

The following shows the evolution of the weight distribution over the 50 Epochs. It is interesting to note how at Epoch 0, the weights are initially close to normally distributed. As the Epochs approach 50, the weights are tuned and follow a specific distribution.

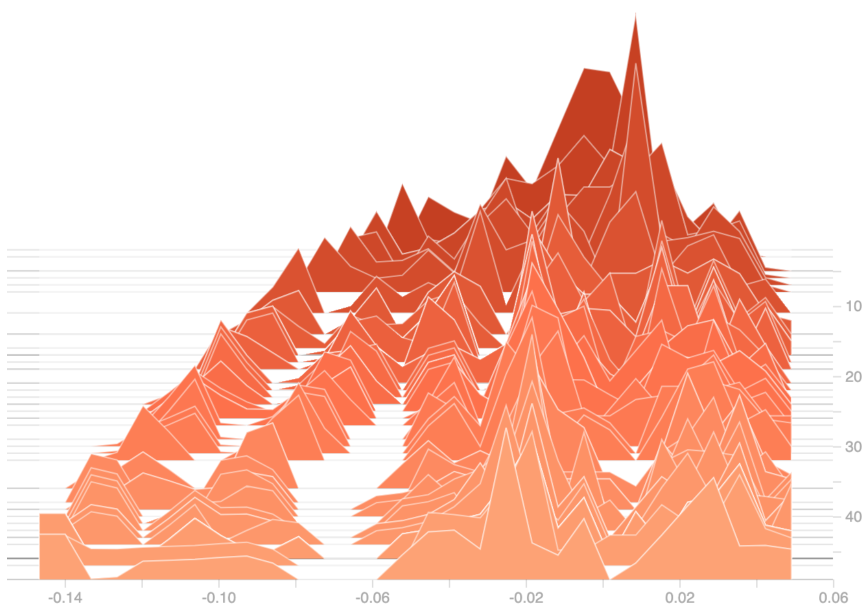

Figure 5.4: Weight Distribution Histogram over 50 Epochs

To be consistent with the evaluation metrics of our baseline models, we tested our model and obtained the confusion matrix shown in Figure (5.5).

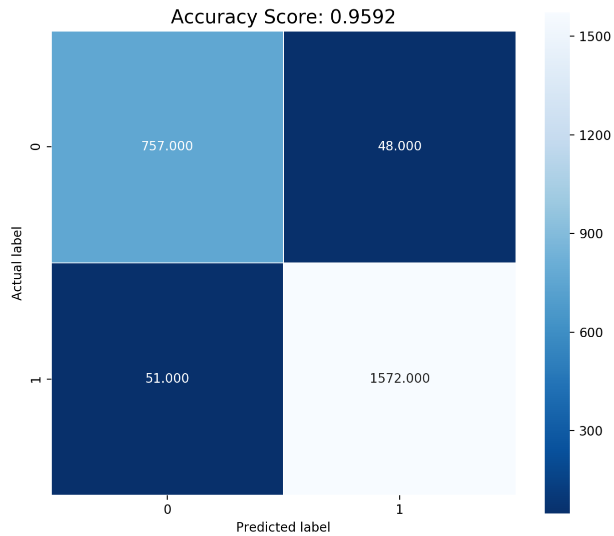

Figure 5.5: Confusion Matrix for CNN

6 Conclusion

The accuracy scores for all models generated in this project using various machine learning methods are listed below.

Model	Testing Accuracy
Logistic Regression	0.7727
Ridge Regression	0.7492
Random Forest	0.7537
CNN	0.9592

The trained Convolutional Neural Network yielded the highest testing accuracy score, 95.92%. CNN is once again proved to be a good machine learning method for processing image data.

7 Future Work

To further improve our solution in the future, we could consider doing the following:

- Use feature extraction to train the model on selected features rather than all the pixels in one image.
- Expand the problem to reproduce probability distribution for all 37 classes.
- Exploit the invariances in images by randomly rotating, scaling, translating, and reflecting them.

8 References

- Alex Krizhevsky et al., 2012. “ImageNet Classification with Deep Convolutional Neural Networks”
- H. Domínguez Spánchez et al., 2019. “Transfer learning for galaxy morphology from one survey to another”
- Kyle W. Willett et al., 2013. “Galaxy Zoo 2: detailed morphological classifications for 304,122 galaxies from the Sloan Digital Sky Survey”
- Sander Dieleman, 2014. "My Solution for the Galaxy Zoo Challenge"