ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 5, Issue 6 , June 2018

Classification of Galaxy Morphological Image Based on Convolutional Neural Network

Wahyono, Muhammad Arif Rahman, and Azhari SN

Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta, Indonesia
International Undergraduate Program in Computer Science, Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta, Indonesia
Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta, Indonesia

ABSTRACT: Astronomers use the term 'morphology' to refer to the structural properties of galaxies. The formation of galaxy involves a complex combination of effects, namely, radioactive cooling, star formation, merging of foreign celestial bodies, and etc.By classifying galaxies into different categories, scientists can build a deeper understanding of they form and evolve and even make an estimation of the amount of time that had passed since the 'Big Bang' until the present. This research aims to classify images of different types of galaxies into three more general categories: Elliptical, Spiral and Irregular. The research is based on convolutional neural network, specially using inception. The total number of images that were used in this research were 206 images of elliptical galaxies, 320 images of spiral galaxies and 184 images of irregular galaxies. The test result showed that images that went through image processing showed a rather poor testing accuracy compared to not using any form of image processing. The best testing accuracy that this research obtained was 78.3%.

KEYWORDS: image processing, convolutional neural network, elliptical galaxy, spiral galaxy, irregular galaxy

I.INTRODUCTION

A galaxy is a gigantic collection of interstellar dust, gas, stellar remnant, stars along with their own solar systems. All held together by the force of gravity. Earth is located in a galaxy called the Milky Way. The Milky Way is a spiral shaped galaxy that has a diameter between 100,000 and 180,000 light years across. We used to think that our galaxy contained all the stars in the universe until in 1920, observations by Edwin Hubble showed that Milky Way is just one of many galaxies in the universe and that each galaxy contains billions or even trillions of stars within it.

There are so many galaxies out there in the universe that there could be as many as tens of billions of undiscovered galaxies and we have only discovered a fraction of that. In recent years, with numerous digital sky surveys across a wide range of wavelengths, astronomy has become an immensely data-rich field. For example, the Sloan Digital Sky Survey will produce more than 50,000,000 images of galaxies in the near future. Studying the morphology of galaxies is one of the most important aspects of answering many of the questions that mankind does not know the answer to yet and that is the creation of the universe. By classifying galaxies into different groups in terms of their structure appearance, scientists will be able to understand the origin and formation of galaxies as well as the evolution process of the universe. Galaxy morphological classification on a large-scale database is important to help astronomers reduce classification errors and to help them produce collections of statistical and observational purposes as well as discovering the mystery of nature at large (de la Calleja & Fuentes, 2004).

Galaxy morphological classification is a system used by astronomers today to divide galaxies into groups based on their visual appearance. In this thesis, galaxies will be divided into three categories based on their visual appearance, Elliptical, Spiral and Irregular. Two of these three types of galaxies are further divided and classified into a system that is now known as the tuning fork diagram, a model of classification of galaxies developed by Edwin Hubble.

With the help of space telescopes that are much more powerful than our eyesight, astronomers are able to look into time and space as far as billions of light years away from Earth and explore millions of galaxies far away from our own.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 5, Issue 6 , June 2018

This research will utilize a deep convolutional neural network called the Inception module, which uses a simple global average pooling instead of fully connected layer to significantly reduce the number of parameters used during the training process (Rahman, 2018). Deep learning combined with convolutional neural network has achieved significant results and a huge improvement in visual detection and recognition with a lot of categories. Raw data images can be trained in deep learning without the need of expert knowledge for optimization of segmentation or feature design.

II. RESEARCH SCOPE

The purpose of this thesis is to achieve a high accuracy in terms of galaxy morphological classification using convolutional neural network algorithm called Inception to classify hundreds of images collected from Google search and divide them into three categories: Elliptical, Spiral and Irregular galaxies, as shown in Figure 1.The dataset will consist of three types of images of galaxies found from Google search using Fatkun Batch Image Download. Specifically, 206 images of elliptical galaxies, 320 images of spiral galaxies and 184 images of irregular galaxies are used in this research.

Figure 1. Three classes of galaxy morphological. From left to right: Elliptical Shaped Galaxy, Spiral Shaped Galaxy and Irregular Shaped Galaxy (en.Wikipedia.org, 2006)

III. RELATED WORKS

Machine learning is one of the most common application of artificial intelligence that is used to solve the problem of galaxy morphology classification. An example machine learning usage to classify galaxy morphology was applied by Gauthier et. al. (2016) where they used both supervised and unsupervised methods to study the Galaxy Zoo dataset of 61,578 pre-classified galaxies. They began their experiment using image pre-processing and followed that up by multi-class classification to classify galaxies as either spiral, elliptical, round, disk or others and found out that the algorithm that produced the best accuracy was random forest, with a classification accuracy of 67%. In addition, they used regression to predict the probabilities of galaxies associated with each class and achieved 94% accuracy. They concluded that the variation of galaxy images are correlated with brightness and eccentricity.

Convolutional neural networks had often been used as the ‘go-to’ method for classifying galaxy morphology images. Recently in 2017, Joseph H. Murrugara LL and Nina S. T. Hirata from Department of Computer Science in Institute of Mathematics and Statistics of University of Sao Paulo, Brazil produced a paper that used convolutional neural network as a method to classify images of galaxies into two categories: elliptical and spiral and achieved a classification accuracy of around 90-91%. Although they achieved a very high result of accuracy, they only categorized the images of galaxies into two categories and did not include ‘Irregular’ shaped galaxies in their result (Murrugara& Hirata, 2017).

In 2003, Jorge de la Calleja and Olac Fuentes presented a method for performing automated morphological galaxy classification using two machine learning algorithms: Locally Weighted Regression and Artificial Neural Network with accuracies of 95.11% and 90.36 % respectively by using the 310 images of galaxies that they obtained from NGC catalogue on the web page of the Astronomical Society of the Pacific, and their classification was taken from the interactive NGC catalogue online at www.seds.org. They also concluded that the Locally Weighted Regression algorithm worked faster and obtained better results than Artificial Neural Network (de la Calleja & Fuentes, 2004).

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 5, Issue 6 , June 2018

In another experiment that was ran by Nour Eldeen M. Khalifa, Mohamed Hamed N. Taha, Aboul Ella Hassanien, and I. M. Selim trained over 1356 images collected from EFIGI catalogue and achieved 97% accuracy in testing accuracy, which outperformed all the other experiments that are used in this literature review. To measure the accuracy of the proposed architecture for classifying galaxies types based on deep convolutional neural networks, five different runs were done, and the median accuracy was calculated (Khalifa et al, 2017).

As part of Stanford's Autumn 2007 Machine Learning final projects, Kasivajhula et al, (2007) compared the performance of three galaxy morphological classification methods of which used the following machine learning algorithms: Support Vector Machines (SVM), Random Forests (RF) and Naïve Bayes (NB). They used an architecture consisting of three stages of processes implemented in Matlab. The first stage was the Image Preprocessing phase where each image is individually scaled, rotated, cropped and centered to appear uniform for more accurate feature extraction. The next stage was the Feature Extraction stage where they measured 6 quantities which they called morphic features of the images and then compressed them using Principal Component Analysis (PCA). Finally, in the Classification stage, they trained and predicted classifications with the features collected from the previous stages using Support Vector Machines (SVM), Random Forest (RF), and Naïve Bayes (NB) machine learning classifiers. Unfortunately, their experiment did not achieve a degree of accuracy as high as the other experiments that are mentioned in this literature review.

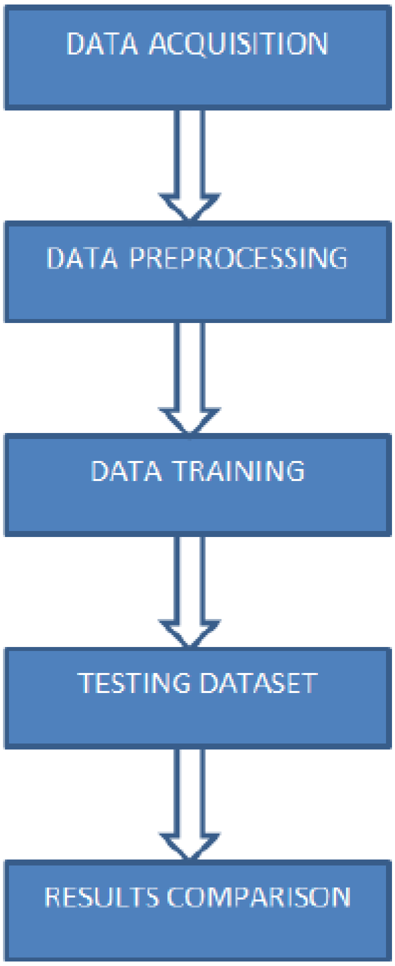

Figure 2. Diagram representation of the steps that will be followed in this research.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 5, Issue 6 , June 2018

IV. METHODOLOGY

The first step of this research will be the gathering of information about galaxy morphological classification methods from previous works related to the subject. This step is done so that further understanding about the proposed topic can be achieved. All the information and knowledge on this paper will be collected from books, journals, scientific paper, article and website that possess subjects related to the topic of this paper.

As can be seen in Figure 2, the steps that will be taken in this research are described below:

1. Data Acquisition: The dataset that will be used in this research are collected manually using the Fatkun Batch Download Image browser extension.
2. Data Pre-processing: Once the data had been collected, they will then undergo three different image processing methods to create invariant images of themselves before they are then fed into the neural network.
3. Data Training: After the raw data had undergone the three different image processing methods, they are then trained using a Tensor Flow library called the Inception module via the Docker Toolbox.
4. Testing Dataset: Once the training process had finished, the testing code would then be implemented to test each image individually to see which category of galaxy they most likely belong to.
5. Results Comparison: The evaluation accuracy that was achieved during the training process will then be compiled into a table and compared to similar researches that involved galaxy morphology classification to see which algorithms is best suited for the particular problem.

The galaxy images downloaded from Google will go through image pre-processing stage to create invariant images to colour, position, orientation and size. Once the pre-processing stage is done, all the images are inputted into the neural network starting from the convolution layer, where the features of the images are extracted. Before the images move on to the classification process, they will all be processed in the pooling layer, where over-fitting is controlled, and computation cost is reduced. The fully connected layer in the classification process is removed due to the Inception module being different from other convolutional neural network algorithms.

Data Acquisition

No coding was done at this stage of the experiment because the images were collected manually (NASA, 2003). Some images might contain noises such as texts, multiple number of galaxies or other subjects that could not be helped to be removed during the image processing stage. This could have played a big factor towards the accuracy of the evaluation accuracy. Figure 3 shows example of data image of elliptical galaxy morphological data.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 5, Issue 6 , June 2018

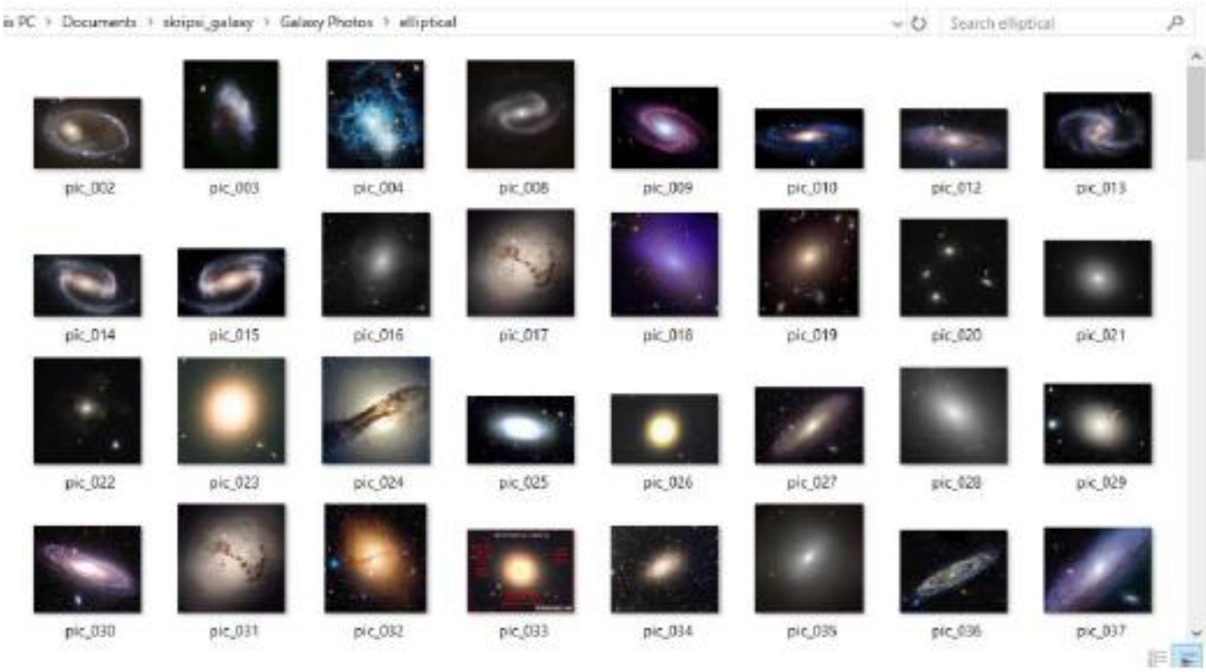

Figure 3. The elliptical galaxy images before image processing

Image Pre-Processing

The data set consists of hundreds of images of galaxies downloaded from Google using a browser extension called Fatkun Batch Image Download. A significant amount of pre-processing is necessary before machine learning algorithm can be applied to the images. The collected dataset is sometimes of different colours, sizes, positioning, and etc. This is where image processing comes in to play, the purpose of image processing will be to create invariant images that are all equal in terms of size and colours among other things before the images are fed to the neural network. The image processing methods that will be applied to the dataset include Canny Edge detection, Histogram Equalization, and Median Filtering (Gonzales, 2008).

Convolutional Neural Network

Convolutional Neural Networks are a compilation of algorithms that are used to identify faces, individuals, street signs, vehicles and many other aspects of visual data. The efficacy of convolutional neural networks in image recognition is one of the main reasons why the world has woken up to the efficiency of deep learning. They are bringing major advancements in computer vision, which has real life applications such as self-driving cars, robots, drones, biometric securities among other things. (Nicholson, 2018).

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 5, Issue 6 , June 2018

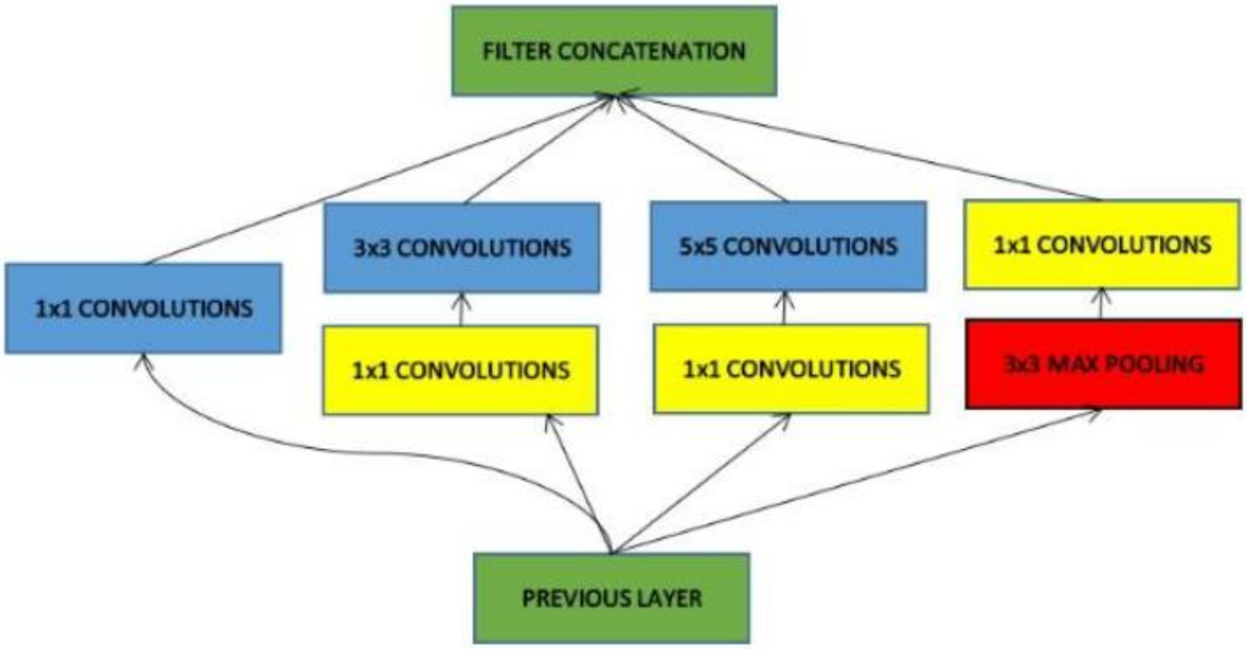

Figure 4. Using 1x1 convolutions to reduce dimensionality in an Inception Module

This research will utilize a deep convolutional neural network called the Inception module, which uses a simple global average pooling instead of fully connected layer to significantly reduce the number of parameters used during the training process.The Inception module is largely more efficient than other image classifying because it is more memory and time efficient. The most important feature of Inception that other modules do not have is the “bottleneck layer”. This bottleneck layer helps in massive reduction of computation requirement.

Inception module different from other convolutional neural network is that it does not have a fully connected layer at the end of the network, instead the module has a simple global average pooling which averages out the channel values across the 2d feature map, after the last convolutional layer. This replacement of fully connected layer drastically reduces the number of parameters that have to be executed and allows the network to contain a large width and depth without affecting the accuracy of the test. (Cv-tricks.com, 2018). In general, as shown in Figure 4, the Inception module is basically consisted of:

- An 299x299x3 input representing a visual field of 299 pixels and 3 color (RGB) channels
- Five vanilla convolution layers, with a few interspersed max-pooling operations
- Successive stacks of “Inception Modules”
- A SoftMax output layer at the end (logits) and at an intermediate output layer (aux_logits) just after the mixed 17x17x768e layer

Overall, these approaches drastically reduce the computational cost of Inception relative to other convolutional neural network architectures.(Chung, 2017).

V. EXPERIMENTAL RESULTS

Once all of the original images have been processed by the three image processing methods, they are then fed to the Inception.v3 module. Table 1 shows the training results of galaxy morphological classification with and without image processing steps.

IJARSET

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 5, Issue 6 , June 2018

Table 1. The results of each image processing methods using epoch ranging from 50-100

Number of epoch	Image Processing Method			
	No Image Processing	Histogram Equalization	Median Filtering	Canny Edge
50	76.7%	53.7%	52.2%	51.6%
60	66.7%	59.7%	56.5%	54.7%
70	68.3%	59.7%	56.5%	56.2%
80	78.3%	55.2%	56.5%	54.7%
90	78.3%	56.7%	56.5%	50.0%
100	76.7%	58.2%	59.4%	51.6%
Average	74.2%	57.2%	56.3%	53.1%

Overall, the inception module gave a better result towards the unprocessed images. This could be caused by the fact that some of the image processing methods might have tampered with the quality of the images. This means that instead of enhancing features, the image processing methods could have resulted in loss of features or even an increase in noise. All of the images that went through the network using each of the image processing methods showed rather poor results, having an average of roughly fifty five percent. The histogram equalization method showed the best result out of the three, this is because this particular method is specially designed to process greyscale images. Canny edge performed the poorestdue to its main purpose being to find the strongest features of an image, making the image clearer was not the main purpose of the canny edge image processing.

When the inception module was applied to the images in the dataset that did not go through any of the image processing stages, it was found that 80-90 epochs showed the most optimum accuracy and as it went past that, the accuracy decreased, this could mean that any epochs before that led to under fitting of the results and any epochs above that could have led to overfitting.

Histogram Equalization found its best accuracy at around 60-70 epochs, so this means, any epochs above that, would only lead to overfitting and decrease in accuracy of the result for this particular image processing method. Median Filtering found a constant increase in testing accuracy as the number of epochs increases. This could mean that this image processing method works best when more number of times the weight are changed in the neural network and when the dataset is passed forward and backward more frequently. It showed that even at 100 epochs, the testing accuracy could still increase and re-testing the dataset using this image processing at a higher number of epochs could even lead to an even further increase of testing accuracy. Canny Edge detection showed the poorest testing accuracy at an average of 53.1%, this could mean that no matter how many epochs are applied to the dataset using this image processing, it would still show rather poor accuracy. This could also be why this image processing method was not the best image processing to use for this particular dataset as it actually had removed some of the most important features that distinguishes one image of galaxy to another.

The highest accuracy that this research achieved was 78.3%, significantly less than most of similar researches that have been done before. This does not mean that the methods of image processing were weak or that the inception module was not on par with other convolutional neural network algorithms in terms of performance. Many factors including the quality of images, the amount of noise that the image originally possessed, and the resizing factor could be the cause of why this research performed poorer than previous similar researches.

One of the biggest factors that could have affected the evaluation accuracy of this experiment might be the colour. Leaving the colour of the images the way they were instead of converting them to grayscale could have resulted in a better evaluation accuracy. Another factor that could have affected the evaluation accuracy of this experiment might have been the resizing factor. The original dataset was converted to 128x128 in the image processing step in this thesis. This size reduction might have been too small and might have impacted the result negatively, therefore, resizing the images to a larger size such as 256x256 might yield better results than 128x128.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 5, Issue 6 , June 2018

Out of all the previous experiments above, M.Khalifa and his colleagues obtained the highest testing accuracy. The dataset that they used were from EFIGI catalogue which consisted more than 13,000 images of galaxies from Hubble telescope, Sloan Digital Sky Survey, NASA Extragalactic Database among a few others. This catalogue consisted excellent quality images of galaxies that were absolutely a class above the images that were used as this researches dataset. Their experiment was run using MATLAB and was CPU specific. To achieve the highest accuracy possible, they used a server with Intel Xeon E5-2620 processor (2 GHz) and 96 GB Ram, which definitely helped out a lot in the computation.

VI.CONCLUSION AND FUTURE WORK

After the research has been concluded, there are a few outcomes that can be taken from the use of different methods of image processing. Those outcomes are as follows:

1. Increasing the number of epoch did not always improve the accuracy of the experiment
2. Histogram equalization obtained the highest training accuracy out of the three image pre-processing methods because that image processing is tailor made for grayscaleimages.
3. The inception module was able to achieve the highest accuracy without any image pre-processing involved. This could potentially mean that this convolutional neural network algorithm is more suited to coloured images than grayscale images.

These are some suggestions that could be taken into consideration for more successful future works.

1. To get a higher accuracy and reliability of the code, more images could be added.
2. Future works could use another form of convolutional neural network such as Residual Networks, Alex Net and VGG to test which methods work more efficiently and which methods put the least or most amount of computation.
3. Different types of Images and more classification categories could also be added to test the reliability of the implemented code.
4. The size and colour of the image could be changed.

REFERENCES

[1] Rahman, M. A. (2018). Galaxy Morphological Classification Using Image Processing and Convolutional Neural Network, Undergraduate Thesis, Submitted to Undergraduate Program in Computer Science, Universitas Gadjah Mada.

[2] Calleja, J. D., & Fuentes, O. (2004). Machine learning and image analysis for morphological galaxy classification. Monthly Notices of the Royal Astronomical Society, 349(1), 87-93.

[3] Gauthier et al (2016) Galaxy Morphology Classification. Available:http://cs229.stanford.edu/proj2016/report/GauthierJainNoordeh-GalaxyMorphology-report.pdf. Last accessed 23th Nov 2017.

[4] H.Murrugarra LL. and Nina S. T. Hirata. (2017). Galaxy image classification. Available: http://sibgrapi2017.ic.uff.br/e-proceedings/assets/papers/WIP/WIP8.pdf. Last accessed 23th Nov 2017.

[5] Khalifa et al (2017). Deep Galaxy: Classification of Galaxies based on Deep Convolutional Neural Networks arXivPrepr. arXiv1709.02245,.

[6] Kasivajhula et al (2007) Morphological Galaxy Classification Using Machine Learning Available: http://cs229.stanford.edu/proj2007/KasivajhulaRaghavanShah-MorphologicalGalaxyClassification.pdf [2007]

[7] Gonzalez, R. and Woods, R. (2008). Digital Image Processing. 3rd ed. [ebook] Available at: https://www.math.uci.edu/icamp/courses/math77c/demos/hist_eq.pdf [Accessed 5 Jun. 2018].

[8] Chris V. Nicholson, S. (2018). A Beginner's Guide to Deep Convolutional Neural Networks (CNNs) - Deeplearning4j: Open-source, Distributed Deep Learning for the JVM. [online] Deeplearning4j.org. Available at: https://deeplearning4j.org/convolutionalnetwork [Accessed 4 Jun. 2018].

[9] Chung, W. (2017). We Need to Go Deeper: A Practical Guide to Tensorflow and Inception. [online] Medium. Available at: https://medium.com/initialized-capital/we-need-to-go-deeper-apractical- guide-to-tensorflow-and-inception-50e66281804f [Accessed 5 Jun. 2018].

[10] Cv-tricks.com. (2018). ResNet, AlexNet, VGGNet, Inception: Understanding various architectures of Convolutional Networks – CV-Tricks.com. [online] Available at: http://cvtricks. com/cnn/understand-resnet-alexnet-vgg-inception/ [Accessed 4 Jun. 2018].

[11] NASA, ESA, and The Hubble Heritage Team (STScI/AURA), (2003), NGC 1427A [ONLINE]. Available at: http://hubblesite.org/newscenter/archive/releases/2005/09/image/a/ [Accessed 2 January 2018].