Summer Internship Project Report on

CONVOLUTED COSMOS : AUTOMATIC CLASSIFICATION OF GALAXY IMAGES USING DEEP LEARNING

LeadingIndia.ai

Submitted To:
Dr. Suneet K Gupta
Asst. Professor
Computer Science & Engg.
Bennett University
Dr. Mohit Agarwal
Ph.D. Scholar
Computer Science & Engg.
Bennett University

Submitted By:
Team Number - 4
Diganta Misra,
Kalinga Institute of Industrial Technology Bhubaeswar, India
Akshat Aman,
Bhilai Institute of Technology, Durg
Smriti Bansal,
ABES Engineering College, Ghaziabad
Meera Saseendran T,
Vidya Academy of Science & Technology, Thrissur

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING
SCHOOL OF ENGINEERING AND APPLIED SCIENCES
BENNETT UNIVERSITY
GREATER NOIDA
UTTAR PRADESH

Chapter 1

ACKNOWLEDGEMENT

No effort in this world is a solo one, neither is this project. We would like to express our heartfelt gratitude to our mentor Dr. Suneet K Gupta who gave us the glorious opportunity to do this wonderful project on the topic "CONVOLUTED COSMOS : AUTOMATIC CLASSIFICATION OF GALAXY IMAGES USING DEEP LEARNING", which made us undergo rigorous research that eventually enhanced our knowledge and understanding.

We are grateful to Dr. Mohit Agarwal, for his esteemed supervision, support and constant encouragement throughout our project work. Now, we would like to express our gratitude and appreciation towards Dr. Deepak Garg HOD-CSE Department and Dr. Suneet Tuli, Dean-School of Engineering and Applied Sciences Bennett University, and all the faculty members of department of engineering for giving us this opportunity to showcase our talent and build the foundation that plays an essential role in our professional lives.

Contents

Chapter 2

Introduction

Studying the categories and also the properties of galaxies are vital because it offers vital clues concerning the origin and also the development of the universe. The classification of the galaxy has a vital role in finding out the formation of galaxies and analysis of our universe. Galaxy morphological classification is done on large databases of information to help astrophysicists in testing theories and finding new conclusions for explaining the physics of processes governing galaxies,star-formation, and the analysis of universe.

Historically, galaxies classification could be a matter of visually inspecting 2 - dimensional pictures of galaxies and categorizing them as they seem. This classificat ion was thought of a long-run goal for astrophysicists. However, the difficult nature of galaxies and quality of pictures have created the classification of galaxies difficult and not correct.

Galaxy classification system helps astronomers within the method of grouping galaxies per their visual form. The foremost notable being the Hubble sequence is considered one among the foremost used schemes in galaxy morphological classificat ion. The Edwin Powell Hubble sequence was created by Hubble in 1926.

In the past few years, advancements in machine tools and algorithms have began to enable automatic analysis of galaxy morphology. There is many machine learning ways which wont to improve the classification of galaxy pictures. Prior researchers don't deliver the goods satisfying results.

Convolution operation is well-known within the laptop vision and signals process community. The convolutional operation is often employed by standard laptop vision, particularly for noise reduction and edge detection. the thought of a Convoluti onal Neural Network (CNN) isn't recent. In 1998, CNN achieved nice results for written digit recognition. However, they dramatically drop because of memory and hardware constraints, besides the absence of enormous coaching information. They were unable to scale to a lot of larger pictures. With the massive increase within the process power, memory size and also the convenience of powerful GPUs and huge datasets, it absolutely was potential to coach deeper, larger and a lot of advanced models.

Chapter 3

Abstract

In this project, Galaxy Image Classification using a Deep Convolutional Neural Network is presented. The galaxy can be classified based on its features into three main categories, namely: Elliptical, Spiral, and Irregular. The proposed deep galaxy architecture consists of one input convolutional layer having 16 filters, followed by 3 hidden layers, 1 penultimate dense layer and an Output Softmax layer. It is trained over 3232 images for 200 epochs and achieved a testing accuracy 97.38% which outperformed conventional classifiers like Support Vector Machine and previous research contributions in the same domain of Galaxy Image Classification.

3.1 Motivation

In past few decades, the desire of humans to know more about other galaxies has increased and so has their efforts. We want to know the most fundamental 3 questions about our existence, how and why. A part of the answer to our question lies in how the galaxies originated and evolved over time. Different galaxies have varying shapes, sizes, colors and features and to solve the puzzle of formation and evolution of galaxies, we need to understand how we can infer the distribution, location and type of galaxies on the basis of their shapes, size and color. This, in turns requires us to classify the galaxy images based on their shapes, sizes and other features.

In an earlier successful projects, hundreds of thousands of volunteers helped classify shapes of some millions of these images by eye. But with growing data, it became difficult to do this manually any more. So, an initiative was launched to find good automated metrics that could potentially be used to analyze the images of the galaxies and answer these questions. The Research will be helpful for astronomical scientists and cosmologists. It will help to classify huge collection of Galaxy images without manual effort of viewing each image individually.

Chapter 4

Literature Survey

[1] John Kormendy and Ralf Bender explains about S0 galaxies which are intermedia te between E7 (elliptical) and Sa (true spiral). They have a bulge and a disk, but no spiral so differs from both elliptical and spiral galaxies and are called lenticular galaxies. Authors make a parallel classification of S0 galaxies to Sa, Sb, Sc and gives them names S0a, S0b, S0c. This classification is done based on B/T ratio, i.e., bulge divided by total light. This value decreases from a to c for both spiral and lenticular galaxies.

[2] Ronald J. buta et al. proposes a new classification of galaxies by name CVHRS. In this, authors classify galaxies in notations of the form: Sab A Sab galaxy that is closer to Sa than Sb, Sab A Sab galaxy that is closer to Sb than to Sa and so on for other galaxies.

[3] Lior Shamir describes automatic classification of galaxy images into elliptical, spiral or edge on galaxies. Manually classifies images are used to extract image features and given Fisher score. Test images are classifies using Nearest Weighted neighbor using the Fisher score as weights. Author finds automatic classification into elliptical, spiral or edge on was done with 90% accuracy.

[4] Edward J. Kim and Robert J. Brunner describes that most star-galaxy classifiers use reduced information from catalogs, this requires careful feature extract ion and selection. With latest advances in machine learning which use deep CNN allows machine to automatically learn the features directly from images and minimiz es need form human input.

[5] Antecedent Research work in the field of classifying Galaxy images correctly into their corresponding major classes didnt yield satisfactory results. In case of, the contributors relied on Machine learning and Image Analysis techniques, to be more precise, a feed-forward neural network and locally-weighted regression method for classification of the Galaxy Image dataset and the accuracy they achieved was approximated to be about 91%.

[6] Nave Bayes Classifiers and Random Forest algorithms for classification was used with which an accuracy of 91% and 79% correspondingly was obtained.

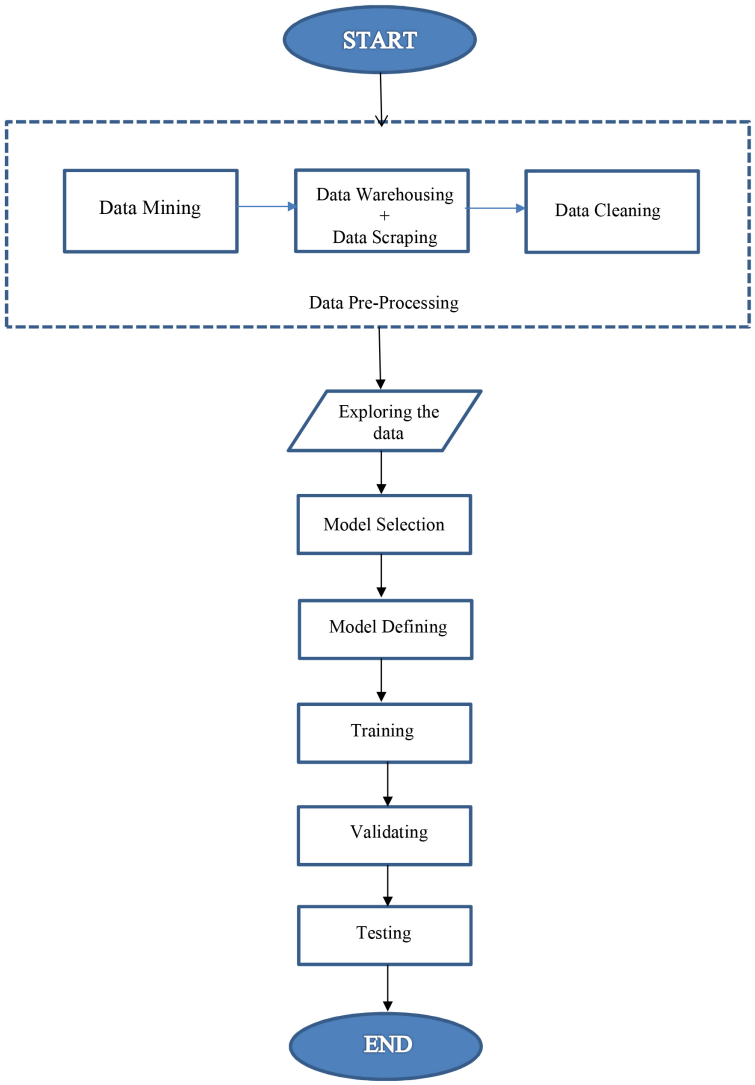
[7] The contributors decided to put Non-Negative Matrix Factorization for the galaxy images, a Supervised Learning technique to use. The accuracy they achieved is stated to be 93%.

[8] In 2017, authors brought forward an innovative automated machine learning technique based on Non-Negative Matrix Factorization, a Supervised Learning techni que again and achieved an accuracy of 92%.

Chapter 5

Project Design

5.1 Data Flow Diagram



Chapter 6

Functionality of the project

In this project, we classified Galaxy Image data into its 3 corresponding major classes - Elliptical type, Spiral type and Irregular type using a Deep Convolutional Neural Network (CNN) architecture.

- **Data Pre-Processing**
  The Dataset containing the Galaxy Images was obtained from Kaggle and NASA Hubble-Space Gallery Websites. The Dataset was categorized into 3 classes with images kept in two main folders: training and validation. The training folder is further subdivided into three subfolders for 3 classes: spiral, elliptical and irregular. Similarly, the validation folder is also having three subfolders with same name.
  The number of images in these folders are listed in table below:

  Table Different classes with number of images for training, validation and testing.

Classes	**Total Images**	**Training Set**	**Validation Set**	**Testing Set**
Spiral	1464	1000	400	64
Elliptical	1464	1000	400	64
Irregular	1686	1232	390	64

  Initially, we had only 11 images for the Irregular class. We used the Augmentor Package of Python to perform Image Augmentation on those 11 images and generated 1615 augmented images.

- **Model Selection**
  We built our Convolutional Neural Network (CNN) model in Python using the Keras framework. The CNN architecture comprised of 1 Input Convolution 2D layer followed by 4 hidden layers, 1 penultimate dense layer and finally 1 output layer. We used a filter size of 3 x 3 in each layer. All the images were resized to 128 x 128 pixels. The batch size used was 64 and was trained for 40 epochs with 10 timestamps per epoch. We used Dropout Regularization after the hidden layers and before the output layer. The detailed architecture diagram is shown in the figure below:

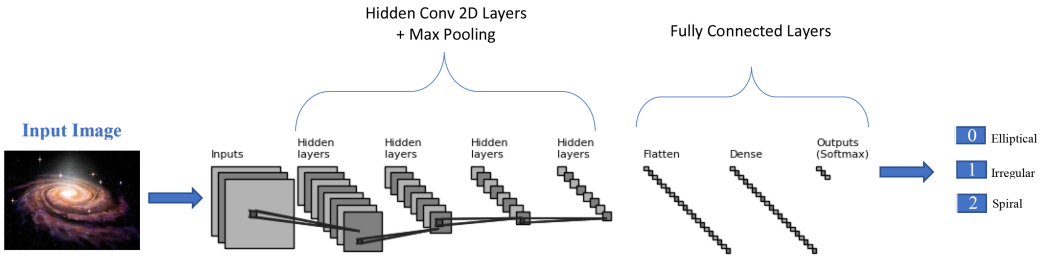


Fig The architecture of Convolution Neural Network for galaxy classification.

- **Training procedure**
  The model was trained on NVIDIA 960MX GPU followed by an intensive training on the NVIDIA DGX 1 Octa Tesla V100 Supercomputer servers using technologies like Putty and WinSCP. On training for 40 epochs, it was observed the training accuracy was at 95.00% with training loss at 15.37% while Validation accuracy was at 94.75% and Validation loss at 15.31%. The Training set containing 3 classes were a total of 3232 images while the Validation set containing the same number of classes contained 1190 images.

  After training the CNN model, we saved the weights of the model in weights.h5 file as we could now perform testing easily as many times we needed using the already trained model weights. This step was also performed as training took nearly 1.5 hours to complete and it was not feasible to train the model every-time for testing. Thus, this time was saved by making the weights.h5 file for testing.

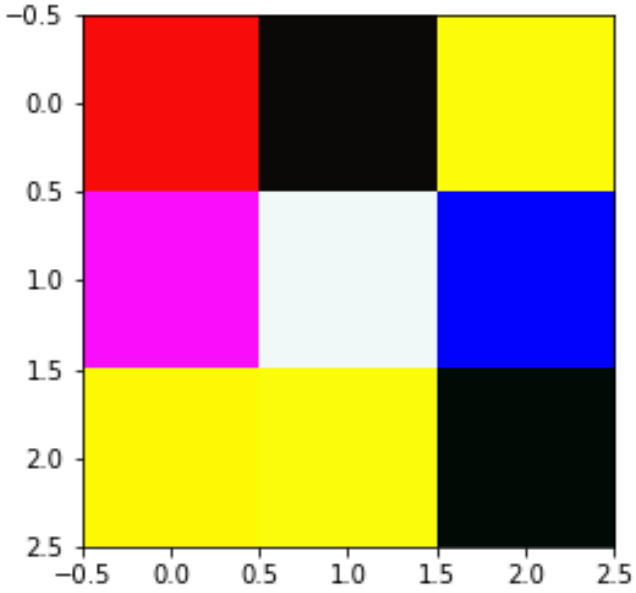


Fig First Layer First Filter Visualization.

Chapter 7

Testing & Implementation

7.1 Testing procedure

We tested our trained Galaxy Classifier CNN architecture model on some new images of the 3 corresponding classes - Irregular, Elliptical and Spiral. After 200 epochs, running on the NVIDIA DGX 1 Octa Tesla V100 Supercomputer servers, we obtained:
Training Accuracy = **99.530%**
Validation Accuracy = **98.480%**
After training and saving the model weights, we tested the model and obtained:
Testing accuracy = **97.398%**
The model accuracy curve is shown below:

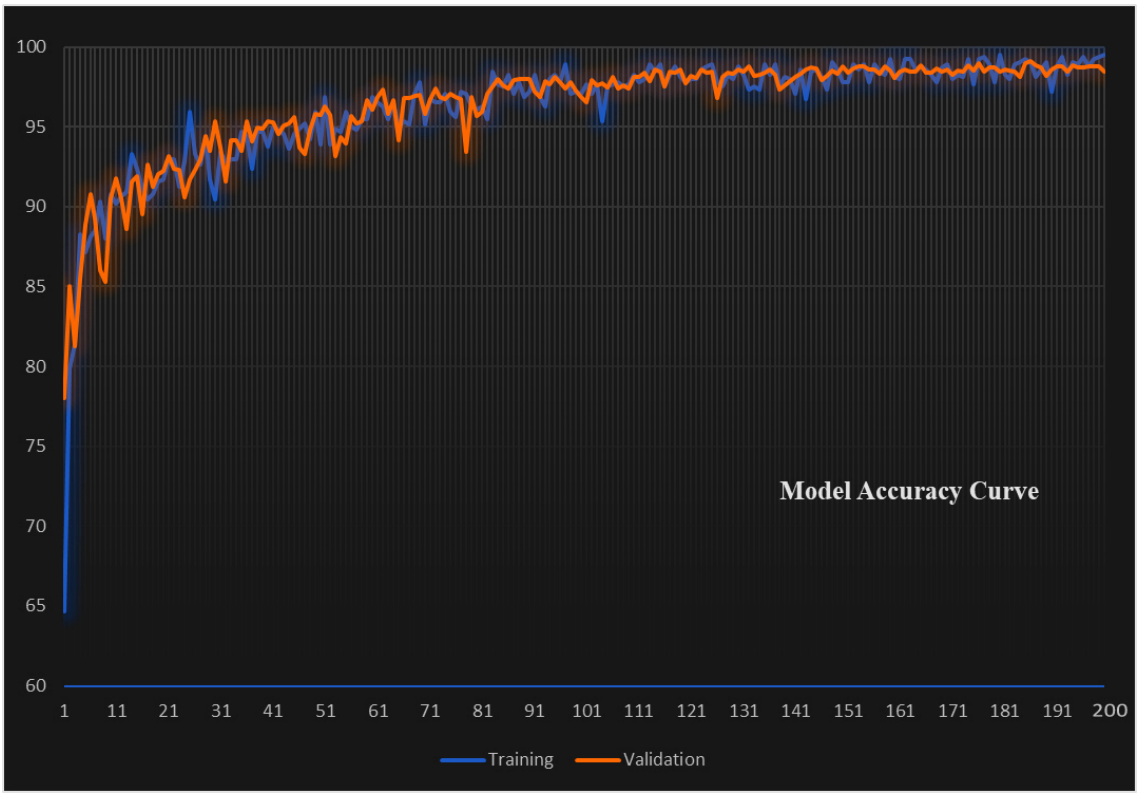


Fig Model Accuracy Curve.

The model loss curve is shown below:

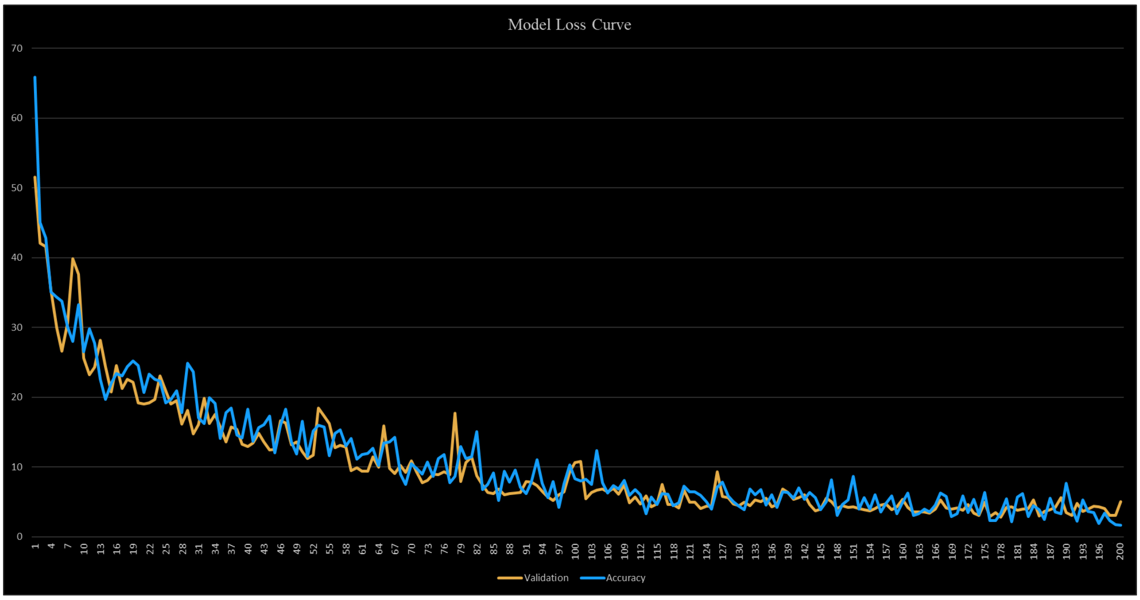


Fig Model Loss Curve.

7.2 Implementation Details

We used a Windows 10 machine having the following specifications: Intel dual-core i5 processor clocked at 2.40 GHz with 8 GB RAM and a dedicated Nvidia GeForce 940M GPU. We also trained the model on the Nvidia DGX 1 (8X Tesla V100) Supercomputer servers having 5120 Nvidia Tensor Cores and a computing power of 960 TeraFlops. The python code was written on Spyder 3.0.0, Jupyter 1.0.0 and Python 3.5.2. using Keras 2.1.3 framework. Many more python packages were needed and used time to time.

Upon testing, the results were obtained in the form of a list with each image having labels as 0, 1 and 2, where 0 stands for Elliptical type, 1 stands for Irregular type and 2 stands for Spiral type.

The sample output for 64 Elliptical images is shown below:
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], dtype=int64)

Similarly, the outputs for all classes are find and accuracy is calculated.

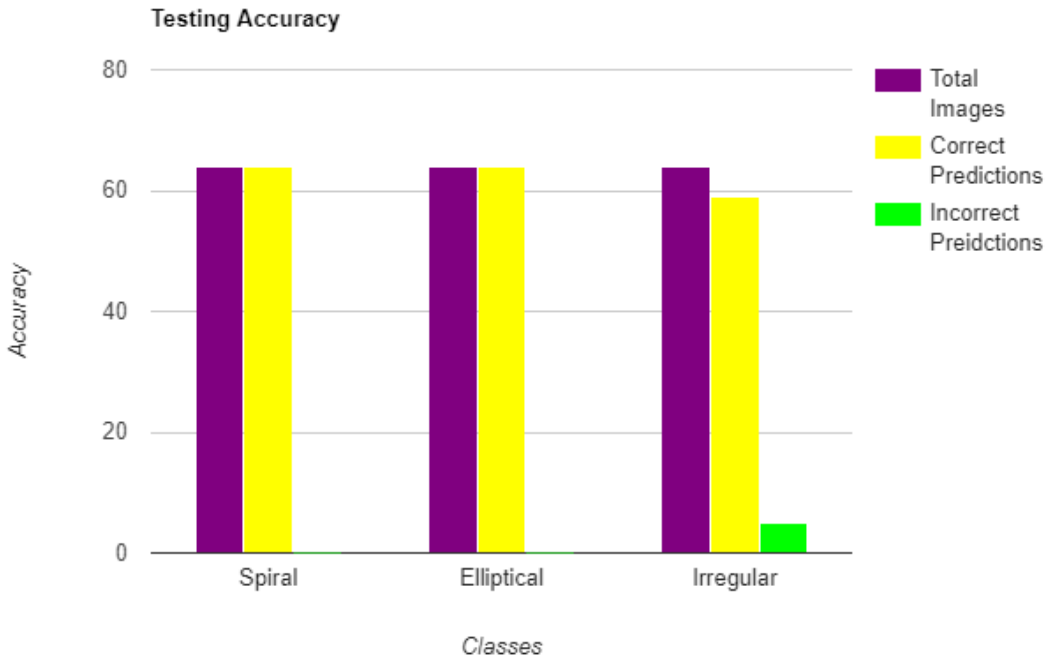


Fig Testing Accuracy Graph for Different Classes.

7.3 Project Development Time Schedule

Our project is pipelined in a very structured manner. The complete project had to be completed within 4 weeks time. So, we divided it into broader sub-areas and started completing them orderly.

- During the first week, we did Data Pre-Processing, which included Data Mining, Data Warehousing, Data Scraping, Data Cleaning etc.
- After the data was processed satisfactorily, we selected our model during the second week and started working on building the CNN model, using Keras framework. Here we worked on selecting the best suitable framework for our model.
- Then the model was trained during the third week. The model was trained on NVIDIA 960MX GPU followed by an intensive training on the NVIDIA DGX 1 Octa Tesla V100 Supercomputer servers, varying the parameters to get the best output. The model took about 1 and a half hour to train on the GPU. After this, the model was extensively tested using various inputs, until satisfactorily results were found.
- Finally, we hyper-tuned the project in the last week. We prepared the visualiza tion layers, accuracy and loss curves and tried every possible way to increase the accuracy. At last, we documented the report and completed the project.

7.4 Learnings & Reflections

We learnt a lot through this project. Right from the data pre-processing to the model testing, each and every phase taught us a new concept and helped boosting our confidence.
Some of the major learnings and reflections can be described as:

- Data is the most important part of any Machine Learning model. The Data Processing part in our project taught us how to manage data, how to classify it, sort the relevant ones and delete the rest, where to store it, augment the data and likewise.
- Selecting which model will be suitable for the project, is also a tedious job. But after this project, we are at least able to make a good assumption for the model selection.
- Which activation function will perform better, how to overcome overfitting and underfitting- all these parameters were learnt during the project.
- Continuously evaluating the model till satisfactory accuracy is achieved- taught us how to remain patient and work on the right aspects.
- We learnt how to make a CNN model more mature and thus more accurate.

Chapter 8

Conclusion

The Research will be helpful for astronomical scientists and cosmologists. It will help to classify huge collection of Galaxy images without manual effort of viewing each image individually. The Research can be fine-tuned for further classification of galaxies into their sub-classes as explained in the Background section. The testing time was reduced to few seconds by saving the CNN weights file and thus it will be working on real time scenarios also.

8.1 Limitations and Future Enhancements

8.1.1 Limitations

- The project is classified only for three broad classes, not for all the sub-classes.
- It has to be run on high speed GPUs in order to get maximum accuracy. Thus, cost of computation is high and a good computation server is needed.
- Data is imbalanced.
- The model takes good time to train.

8.1.2 Future Enhancements

- The model can be fine-tuned more to make it more efficient.
- Data can be made balanced to avoid misclassification.
- Model could be made such that it can be run on normal CPUs, rather than running on GPUs in super computers.
- More accurate model can be prepared in future using even fine parameters or new techniques.

Chapter 9

References

[1] John Kormendy and Ralf Bender.,A revised parallel-sequence morphological classification of galaxies: structure and formation of s0 and spheroidal gakaxies, *The Astrophysical Journal Supplementary Series.*,198(1):2,2011

[2] Ronald J. buta.,Kartik Sheth.,E. Athanassoula., A. Bosma., Johan H. Knapen., Eija Laurikainen., Heikki Salo.,Debra Elmegreen., Luis C. Ho., Dennis Zaritsky,A Classical Morphological Analysis of Galaxies in the Spitzer Survey of Stellar Structure in Galaxies, *The Astrophysical Journal Supplementary Series.*, 217(2):32, 2015

[3] Lior Shamir., Automatic morphological classification of galaxy images, *Monthly Notices of the Royal Astronomical Society*, 399(3):13671372,2009

[4] Edward J. Kim and Robert J. Brunner.,Stargalaxy classification using deep convolutional neural networks, *Monthly Notices of the Royal Astronomical Society*, 464(4), 1: 44634475,2017

[5] Jorge De La Calleja and Olac Fuentes.,Machine learning and image analysis for morphological galaxy classification*Monthly Notices of the Royal Astronomical Society*, 349(1):8793, 2004

[6] Maribel Marin and L. Enrique Sucar and Jesus A. Gonzalez and Raquel Diaz.,A Hierarchical Model for Morphological Galaxy Classification, In *FLAIRS conference*, 2013

[7] I.M.Selim.,Arabi E. Keshk.,Bassant M.El Shourbugy.,Galaxy Image Classifica tion using Non-Negative Matrix Factorization,*International Journal of Computer Applications*, 137(5), 2016

[8] I.M.Selim., Mohamed Abd El Aziz.,automated morphological classification of galaxies using projection gradient nonnegative matrix factorisation algorithm,*Experi mental Astronomy*, 43(2):131-144, 2017