VISVESVARAYA TECHNOLOGICAL UNIVERSITY

"Jnana Sangama", Belagavi , Karnataka, INDIA

A Project Report
on

Analyzing and Processing Astronomical Images using Deep Learning

Submitted in partial fulfillment of the requirement for the award of the degree of

Bachelor of Engineering
in
Computer Science and Engineering

Submitted By

SANDEEP V Y	**1GA17CS134**
SANTOSH K	**1GA17CS137**

Under the Guidance of

Mrs.Snigdha sen
Assistant Professor

Department of Computer Science and Engineering
Accredited by NBA(2019-2022)
GLOBAL ACADEMY OF TECHNOLOGY
Rajarajeshwarinagar, Bengaluru - 560 098

2020 – 2021

GLOBAL ACADEMY OF TECHNOLOGY
Department of Computer Science and Engineering
Accredited by NBA(2019-2022)

CERTIFICATE

Certified that the Project Entitled **"*Analyzing and Processing Astronomical Images using Deep Learning*"** carried out by **SANDEEP V Y**, bearing **USN 1GA17CS134, SANTOSH K**,bearing **USN 1GA17CS137,** bonafide students of Global Academy of Technology, is in partial fulfillment for the award of the **BACHELOR OF ENGINEERING** in **Computer Science and Engineering** from Visvesvaraya Technological University, Belagavi during the year 2020-2021. It is certified that all the corrections/suggestions indicated for Internal Assessment have been incorporated in the report submitted to the department. The Partial Project report has been approved as it satisfies the academic requirements in respect of the project work prescribed for the said Degree.

Mrs.Snigdha sen Assistant professor Dept. of CSE GAT, Bengaluru.	Dr. Srikanta Murthy K Professor & HOD Dept. of CSE GAT, Bengaluru.	Dr. Rana Pratap Reddy Principal GAT, Bengaluru.

GLOBAL ACADEMY OF TECHNOLOGY

Rajarajeshwarinagar, Bengaluru – 560 098

DECLARATION

We, **SANDEEP V Y**, bearing **USN 1GA17CS134, SANTOSH K**,bearing **USN 1GA17CS137,** students of Seventh Semester B.E, Department of Computer Science and Engineering, Global Academy of Technology, Rajarajeshwarinagar Bengaluru, declare that the Project Work entitled "***Analyzing and Processing Astronomical Images using Deep Learning***" has been carried out by us and submitted in partial fulfillment of the course requirements for the award of degree in **Bachelor of Engineering** in **Computer Science and Engineering** from **Visvesvaraya Technological University, Belagavi** during the academic year **2020-2021**.

SANDEEP V Y 1GA17CS134
SANTOSH K 1GA17CS137

Place: Bengaluru
Date: 15/1/2021

ABSTRACT

Distance in space is an important parameter, by knowing the distance, we can find mass,luminosity, star formation rate, metallicity etc. We can find the distance of galaxies using their redshift values.

Currently, astronomical community is experiencing a data deluge. Machine learning and, in particular, deep-learning technologies are increasing in popularity and can deliver a solution to automate complex tasks on large data sets. The abundant photometric data collected from multiple large-scale sky surveys give important opportunities for photometric redshift estimation. However, the low accuracy is a serious issue which still exists in the current photometric redshift estimation methods. Many researchers have contributed in this area, but still there is a significant gap to fill in.

In this work, the challenge of classifying galaxies and deriving redshift values from photometric data is addressed. Existing methods on this task are quite complex and error is high, so to simplify the process of predicting the redshift, to classify astronomical images and to improve the accuracy of existing redshift estimation, we implement new method of estimation by combining SOM - CNN technique.

This project aims to analyze the photometric data collected from multiple large-scale sky surveys and classifies the astronomical images into elliptical, spiral, and lenticular classes through Self Organizing Map (SOM) method to predict their redshift values through Convolutional Neural Networks (CNN) with good accuracy using pretrained weights of various Deep learning architectures like AlexNet, VGG16, ResNet50, InceptionV3, Xception, etc.

Various algorithms and pretrained architectures are implemented on Galaxy image dataset using real-time data augmentation and its performance are evaluated. The accuracy of model in classifying the astronomical images is above 93 % with validation accuracy above 88 % and loss in predicting redshift value is less than 0.01 and is varying with various CNN models. Our experimental results show the significant improvement over existing techniques used.

ACKNOWLEDGEMENT

The satisfaction and the euphoria that accompany the successful completion of any task would be incomplete without the mention of the people who made it possible. The constant guidance of these persons and encouragement provide, crowned our efforts with success and glory. Although it is not possible to thank all the members who helped for the completion of the phase - 1 of project individually, we take this opportunity to express our gratitude to one and all.

We are grateful to management and our institute **GLOBAL ACADEMY OF TECHNOLOGY** with its very ideals and inspiration for having provided us with the facilities, which made this, phase - 1 project a success.

We express our sincere gratitude to **Dr. N. Rana Pratap Reddy,** Principal, Global Academy of Technology for the support and encouragement.

We wish to place on record, our grateful thanks to **Dr. Srikanta Murthy K**, HOD, Department of CSE , Global Academy of Technology, for the constant encouragement provided to us.

We are indebted with a deep sense of gratitude for the constant inspiration, encouragement, timely guidance and valid suggestion given to us by our guide **Mrs.Snigdha sen, Assistant Professor**, Department of CSE, Global Academy of Technology.

We are thankful to all the staff members of the department for providing relevant information and helped in different capacities in carrying out this phase -1 project.

Last, but not least, we owe our debts to our parents, friends and also those who directly or indirectly have helped us to make the phase - 1 project work a success.

SANDEEP V Y **1GA17CS134**

SANTOSH K **1GA17CS137**

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

SRS	Software Requirement Specification
DFD	Data Flow Diagram
CNN	Convolutional Neural Networks
SOM	Self-Organizing Map
DL	Deep Learning

CHAPTER 1

INTRODUCTION

1.1 Definitions

Deep learning can be considered as a subset of machine learning. It is a field that is based on learning and improving on its own by examining computer algorithms. While Machine learning uses simpler concepts, deep learning works with artificial neural networks, which are designed to imitate how humans think and learn. The ability to process large numbers of features makes deep learning very powerful when dealing with unstructured data.

Artificial Neural Networks (ANN) is a series of algorithms that endeavours to recognize underlying relationships in a set of data through a process that mimics the operations of human brain. It works by adjusting “weights” during training, and predicts the output based on these adjusted weights.

Convolution Neural Networks (CNN) are deep learning algorithms that can train large datasets with millions of parameters, in form of 2D images as input and convolve it with filters to produce the desired outputs.

CNN is most often used to analyze visual imagery. They have applications in image and video recognition, recommender systems, image classification, medical image analysis, natural language processing, and financial time series.

The Self-Organizing Map (SOM) method is a new, powerful software tool for the visualization of multi-dimensional data. It converts complex, non-linear statistical relationships among high-dimensional data into simple geometric relationships on a low-dimensional display.

1.2 Project report outline

Redshift estimation is a very challenging task in astronomy and need more manual power to accomplish. Many machine learning models had been created to come up with good accuracy in estimation without manual intervention. Constant improvements in performances had been achieved by adopting and modifying machine learning approaches. The need for precise redshift estimation is increasing due to its importance in cosmology.

In this project, CNN models are built to predict redshift values from photometric galaxy data images and to classify them to their respective classes. We implement new method of estimation by combining SOM - CNN techniques with various CNN architectures to analyze and to predict the redshift values of galaxies.

This report covers the introduction of the project, the system requirements, hardware and software requirements, functional and non-functional requirements, the literature survey, the model architectures used and their results.

CHAPTER 2

REVIEW OF LITERATURE

2.1 System Study

In the visible band, the wavelength of spectrum will increase which looks like moving toward the red side of the band due to the star is flying away. This phenomenon is called ”redshift”. Astronomers use redshift to measure approximate distances to very distant galaxies. The more distant an object, the more it will be redshifted. As the universe expands, the space between galaxies is expanding. Measuring the redshift directly is a time consuming and expensive task as strong spectral features have to be clearly recognized. Therefore, redshifts extracted via photometry based models provide a good alternative.

In recent years, the availability of large synoptic multi-band surveys increased the need of new and more efficient data analysis methods. The astronomical community is currently experiencing a data deluge. Machine learning and, in particular, deep-learning technologies are increasing in popularity and can deliver a solution to automate complex tasks on large data sets. In astronomy, machine learning techniques have been applied to many different uses. Redshift estimation is just one relevant field of application for the statistical methods. Most machine-learning based photometric redshift estimation approaches found in the literature just generate single value estimates.

2.2 Proposed Work

Many researchers used various supervised and unsupervised learning algorithms like nearest neighbor algorithm, neural network method, linear spectral connectivity analysis, Arbor Z boosted decision tree method, random forest to calculate redshifts using photometric attributes on a spectroscopic training set. But, still there is a huge demand for estimating precise redshift value from telescopic survey data.

Our proposed project provides a novel two-stage photometric redshift estimation approach, i.e. the integration of Convolutional Neural Network (CNN) and Self organizing map (SOM), to improve the estimation accuracy. We are also training the model with various CNN architectures like AlexNet, VGG16, ResNet50, InceptionV3, Xception, etc., by setting and tuning the hyper-parameters to get good accuracy in classifying the astronomical images.

2.3 Scope of the Project

The challenge of deriving redshift values from photometric data is a complex and monotonous task to do. Many researchers have contributed in this area, but still there is a significant gap to fill in. Convolutional Neural Network has been used in astronomy widely, such as spectral classification and galaxies classification. However, this method has not been applied to the prediction of photometric redshift. Self organizing map (SOM) is able to learn independently and automatically adjust network parameters and structures according to sample characteristics, which can improve the efficiency of CNN. Selection of feature extraction and classification algorithms plays an important role in this area. Deep Learning with Convolution Neural Networks provides us a combination of feature extraction and classification in a single structure. Our experimental results show the significant improvement over existing techniques used in this area.

Chapter 3

System Requirement Specification

3.1 Functional Requirements

- The ultimate goal of this project is to simplify the process of predicting the redshift and to classify astronomical images with good accuracy.
- While training the model it should report the accuracy of the estimated redshift.
- Predictions on test data should be projected.
- Model should enable uploading photometric test images inorder to check whether it is working properly.

.

3.2 Non Functional Requirements

- The developed model is scalable, the model is able to predict redshift using photometric data.
- Appropriate photometric data should be made available to perform this task.
- All the hardware and software requirements has to be met for estimating.
- The model requires its own (minimum) time to accomplish the task.

3.3 Hardware Requirements

- Intel Core i3 processor and above.
- 4 MB RAM (min).
- 1TB Hard Disk Drive.
- Mouse or other pointing device.
- Keyboard.
- Display device.

3.4 Software Requirements

- Google colaboratory or Jupyter notebook.
- Python 3 and above.
- Tensor flow 2.1 and above.
- Any web browser : Chrome, Firefox, Safari.
- Operating System : Linux or Windows.

Chapter 4

System Design

4.1 Design Overview

System Design is the process of defining architecture, components, and modules. The System Design Document describes the system requirements, operating environment, system and subsystem architecture, files and database design, input formats, output layouts, human-machine interfaces, detailed design, processing logic, and external interfacest. There is some overlap with the description of system analysis, system architecture and system engineering.

System design is therefore the process of defining and developing systems to satisfy specified requirements of the user. Object oriented analysis and design methods are becoming the most widely used methods for computer system design. The UML has become the standard language in object oriented analysis and design. It is widely used for modeling software systems and is increasingly used for high designing non-software systems and organizations.

System design is one of the most important phases of software development process. The purpose of the design is to plan the solution of the problem specified by the requirement documentation. In other words the first step in the solution to the problem is the design of the project.

The design will contain the specification of all the modules, their interaction with other modules and the desired output from each module. The output of the design process is a description of the software architecture.

4.2 System Architecture

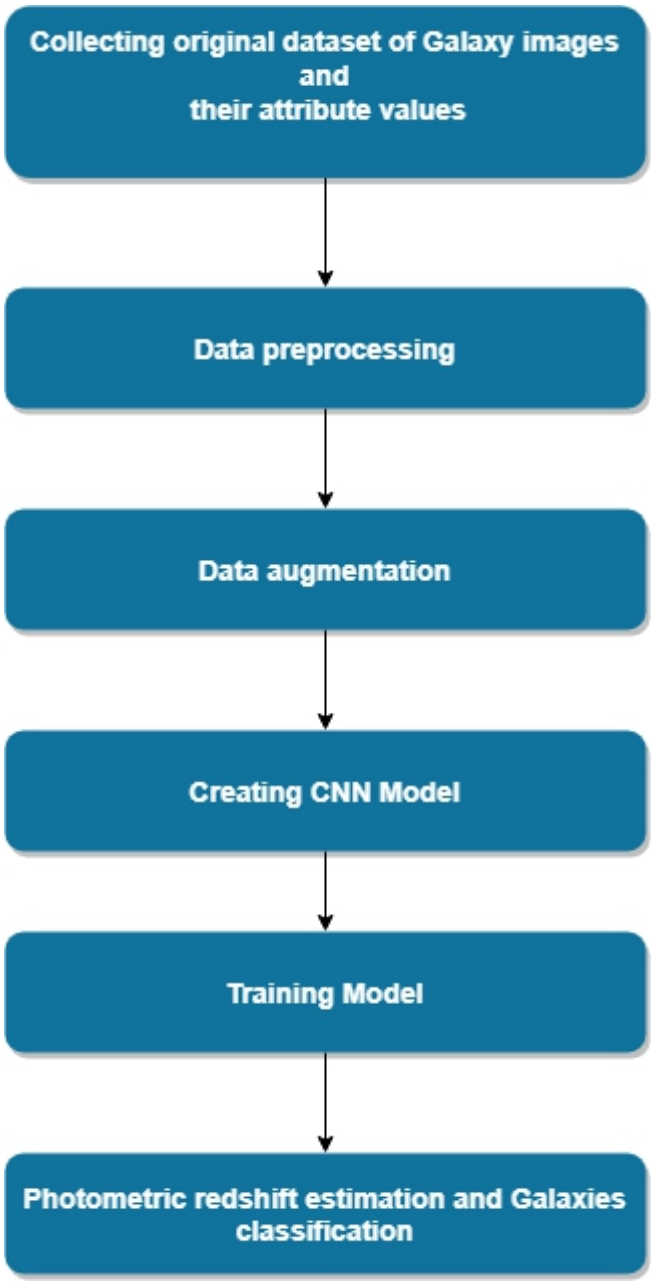

Figure 4.1: System Architecture

The above figure shows a general System Architecture describing the activities performed by this project.

4.3 Data Flow Diagrams

A data flow diagram is the graphical representation of the flow of data through an information system. DFD is very useful in understanding a system and can be efficiently used during analysis.

A DFD shows the flow of data through system. It views a system as a function that transforms the inputs into desired outputs. Any complex systems will not perform this transformation in a single step and data will typically undergo a series of transformations before it becomes the output.

With a data flow diagram, users are able to visualize how the system will operate that the system will accomplish and how the system will be implemented. Old system DFDs can be drawn up and compared with a new system DFD to draw comparisons to implement a more efficient system.

Data flow diagrams can be used to provide the end user with a physical idea of where the data they input, ultimately as an effect upon the structure of the whole system.

4.3.1 Data Flow Diagram - Level 0

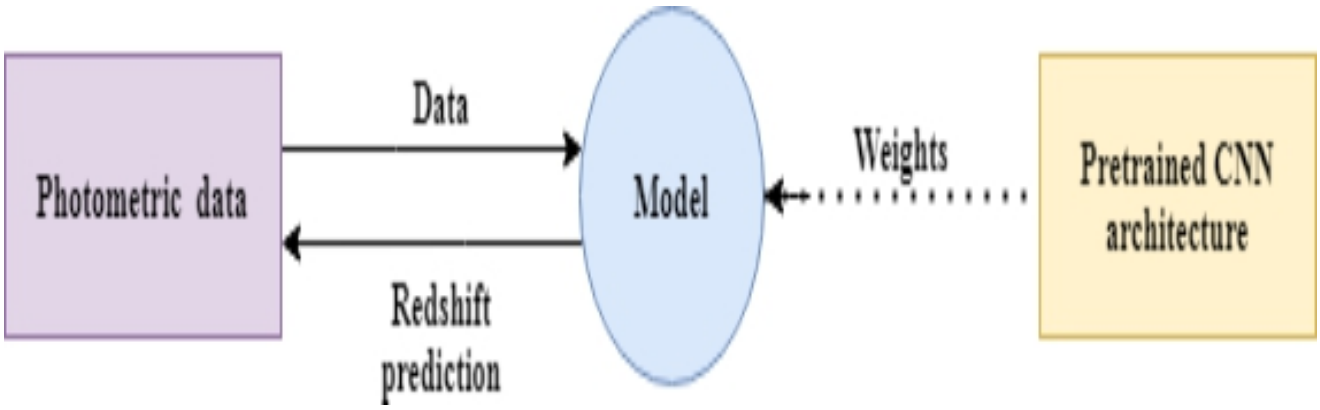

Figure 4.3.1: Data Flow Diagram Level 0

4.3.2 Data Flow Diagram - Level 1

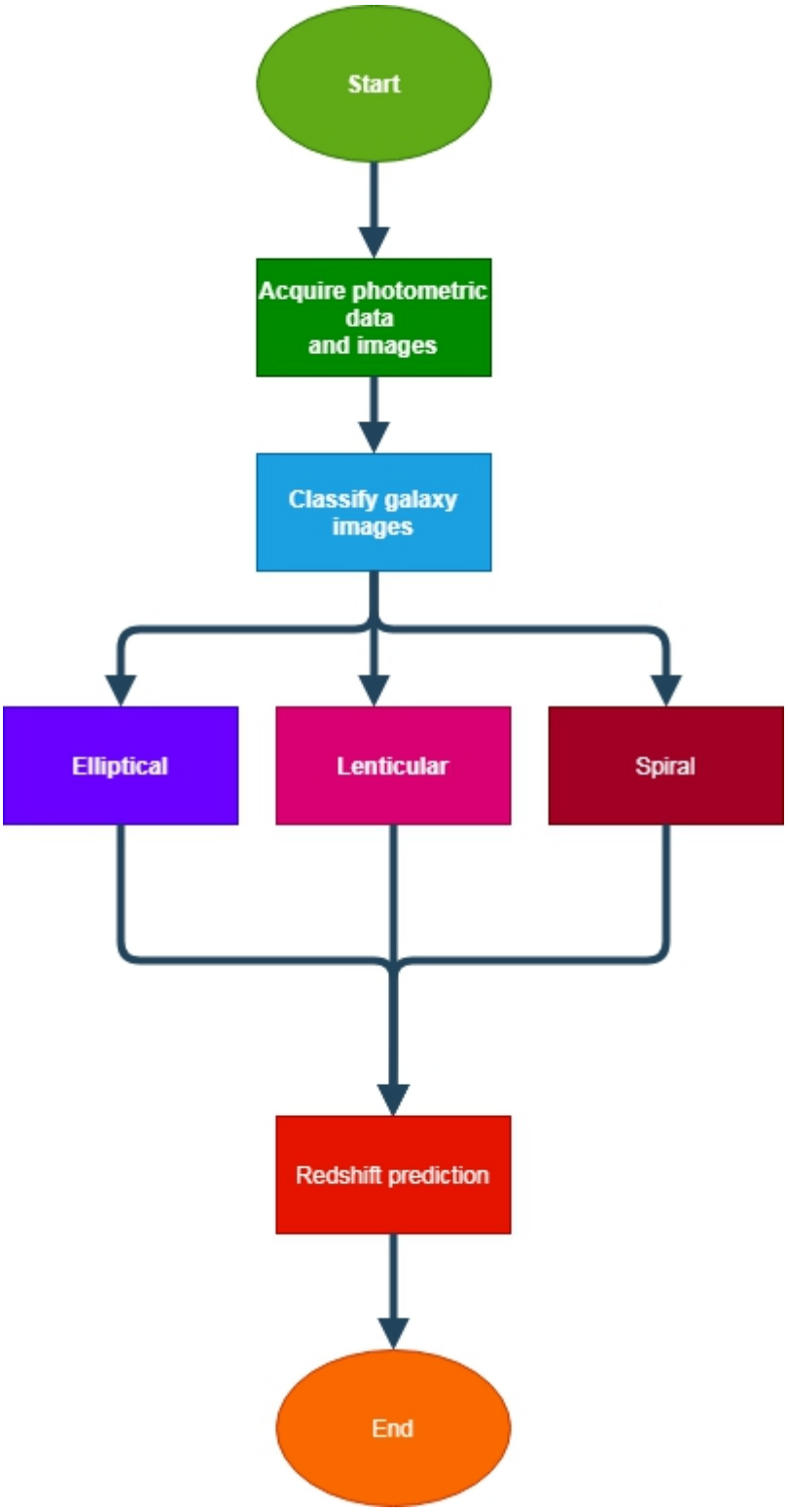

Figure 4.3.2: Data Flow Diagrams Level 1

4.3.3 Data Flow Diagram - Level 2

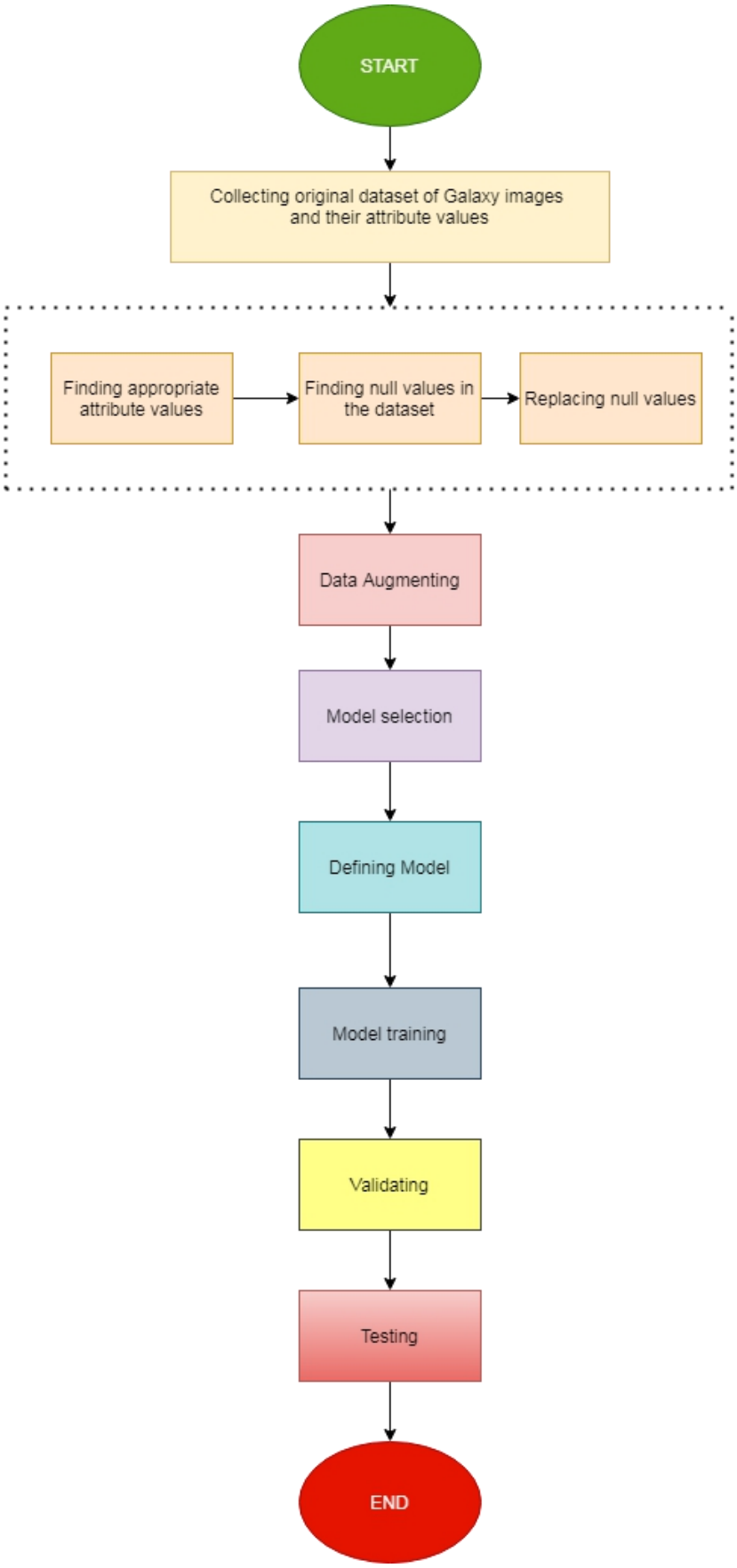

Figure 4.3.3: Data Flow Diagram Level 2

4.4 CNN Architecture Diagrams

CNN architecture is inspired by the organization and functionality of the visual cortex and designed to mimic the connectivity pattern of neurons within the human brain. The neurons within a CNN are split into a three-dimensional structure, with each set of neurons analyzing a small region or feature of the image.

4.4.1 Custom CNN Architecture Diagram - Level 0

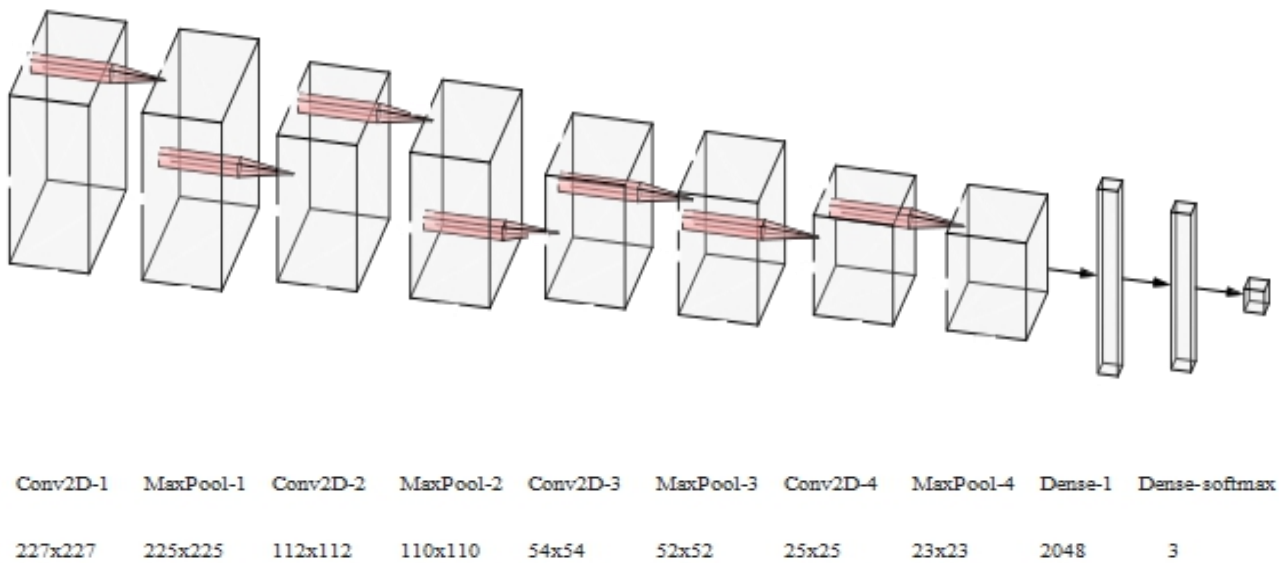

Figure 4.4.1: Custom CNN Architecture Diagram - Level 0

In our custom CNN architecture we used 8 hidden layers (Conv2D + MaxPool) and 2 fully connected layer (Dense+softmax). Flattening and Dropout layers are mounted inbetween hidden and fully connected layers.

Order of placement of these layers are shown accurately in the above diagram.

4.4.2 Custom CNN Architecture Diagram - Level 1

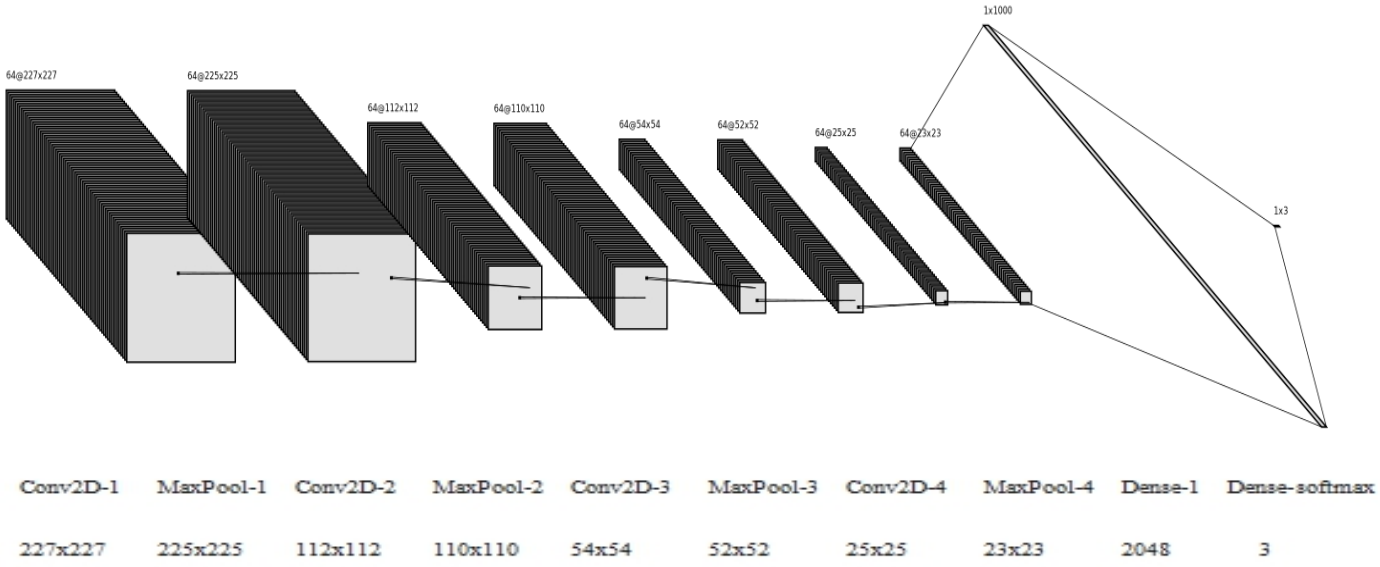

Figure 4.4.2: Custom CNN Architecture Diagram - Level 1

Each layers details are shown in this below table.

Layers	Filters	Input size	Filter size	Parameters	Activation
Conv2D-1	64	227x227	3x3	1792	Relu
MaxPool-1	64	225x225	3x3	0	----------
Conv2D-2	64	112x112	3x3	36928	Relu
MaxPool-2	64	110x110	3x3	0	----------
Conv2D-3	64	54x54	3x3	36928	Relu
MaxPool-3	64	52x52	3x3	0	----------
Conv2D-4	64	25x25	3x3	36928	Relu
MaxPool-4	64	23x23	3x3	0	----------
Dense-1	----------	227x227(7744)	----------	15861760	Relu
Dense-softmax	----------	2048>>3	----------	6147	Softmax

Table 4.4.2: Custom CNN layers details

4.5 Modules

4.5.1 Collecting Photometric Data

Collecting original dataset of galaxy images and their attribute values from multiple large-scale sky surveys give important opportunities for photometric redshift estimation. So, this process is done as the first step in the project.

4.5.2 Data Preprocessing

Noisy data is a meaningless data that can't be interpreted by machines. It can be generated due to faulty data collection, data entry errors etc. So, Data preprocessing is done to transform the raw data in a useful and efficient format.

4.5.3 Data Augmentation

Data augmentation is a strategy that enables practitioners to significantly increase the diversity of data available for training models, without actually collecting new data. Data augmentation techniques such as cropping, padding, and horizontal flipping are commonly used to train large neural networks.

4.5.4 Selecting and Defining CNN Model

Selection of CNN architecture is a very vital part in this project which directly decides the performance of our CNN Model. In our project we created and defined our own custom CNN architecture model and also used various CNN architectures to compare custom CNN model accuracy.

4.5.5 Training CNN Model for Classification and Redshift Prediction

Different CNN models are Trained by tuning and setting various hyperparameters to get good accuracy with different number of epochs. This is the longest phase in our project. Models are trained for classification and redshift prediction of galaxy images.

4.5.6 Validating and Testing

This is the phase where our trained CNN models are validated and tested to estimate their performance. Output of these models are then evaluated and tested against various other CNN architectural models.

CHAPTER 5

RESULTS

In our project phase-1, we have trained the CNN model to classify the galaxy images into Elliptical, Lenticular, and Spiral and to predict their redshift value. The same is evaluated and the results of various CNN models are shown below.

5.1 CNN models trained for galaxy classification

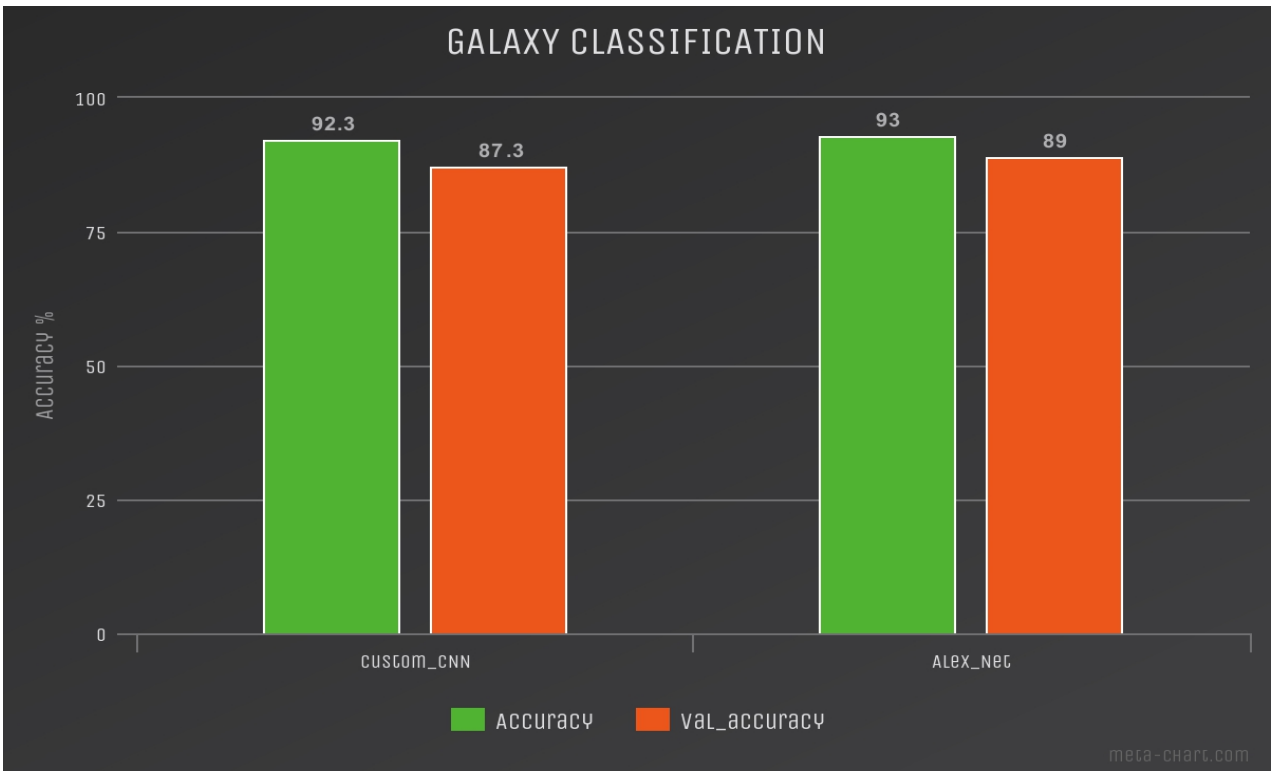

Figure 5.1: Galaxy classification by Custom CNN and AlexNet

Here both our custom CNN and AlexNet models are trained and validated against photometric data. Both models took more than 5 hours to get trained and gave above 90% accuracy with less validation loss.

Below two graphs show the training and validation accuracy/loss respectively for our custom CNN model.

Custom CNN model

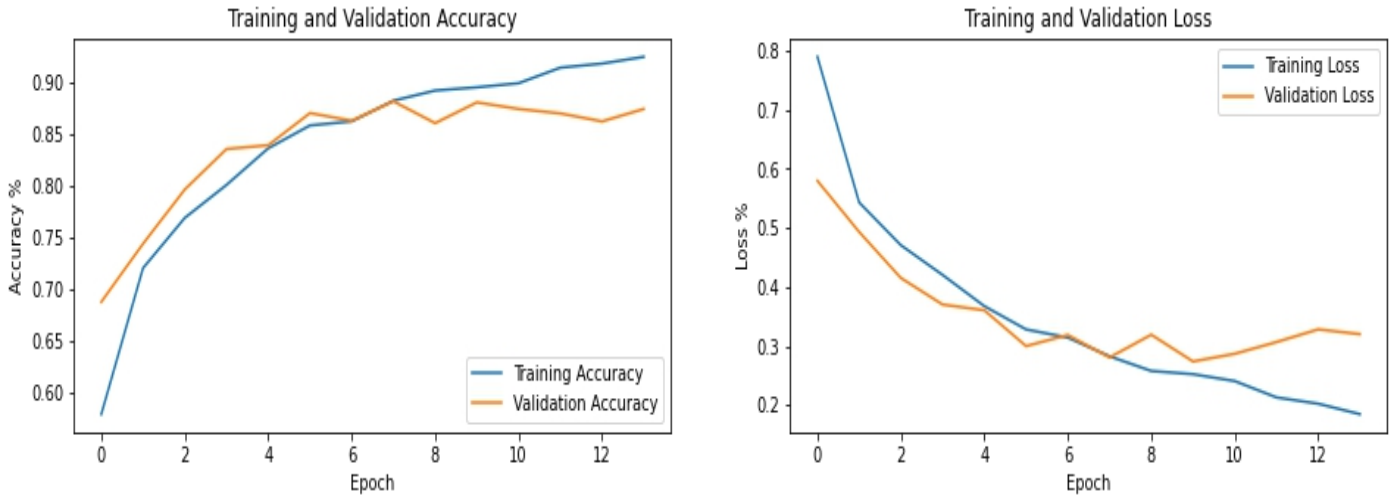

Figure 5.1: Custom CNN model's training and validation

5.2 CNN models trained to predict Redshift values

Architecture	Loss	Validation_Loss	Outliers
Custom_ CNN	0.000167	0.002980	0.012520852861786363
VGG16	0.013038	0.026891	0.012159123986981065
ResNet50	0.012959	0.009488	0.08018843909052954
InceptionV3	0.121746	0.044373	0.055478848792356525
Xception	56.61079	18.498215	1.6760335886666295

Table 5.2 : Various CNN model's performance

From the above table we can conclude that our custom CNN model defined to predict the redshift is doing good than other pretrained model

5.3 Learnings and 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/Deep 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 atleast 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 6

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 project also helps in estimating Redshift value which is really a complex and monotonous task to do. The testing time was reduced to few seconds by saving the CNN weights in file and thus it will be working on real time scenarios also.

The Research can be fine-tuned for further increased accuracy in classification of galaxies and in Redshift prediction by implementing SOM method as discussed earlier. In project phase-2 we will be implementing both SOM - CNN method to improve our current estimation.

BIBILOGRAPHY

[1] 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

[2] Angus H. Wright1, Hendrik Hildebrandt1, Jan Luca van den Busch1 and Catherine Heymans1,2

[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 classificationMonthly 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