Morphological Classification of Galaxies using Artificial Intelligence

Arihant Tiwari (19050)
Department of Physics,
IISER Bhopal, Bhopal, India
arihant19@iiserb.ac.in

Gourav Kumawat (19127)
Department of Physics,
IISER Bhopal, Bhopal, India
gourav19@iiserb.ac.in

Shaunak Padhyegurjar (19213)
Department of Physics,
IISER Bhopal, Bhopal, India
shaunak19@iiserb.ac.in

***Abstract*—In this paper, we present a set of evolutionary and chronologically constrained set of models based on CNN architecture, optimised to n-index classification and adaptable to computational power of the running machine. We focus on defining methods and procedures for morphological classification of galaxies based on the Hubble Classification scheme, using Artificial Intelligence. Previous data cleaning and classification has primarily relied on manual image recognition and assorting, and thus has been unable to cope with the exponentially increasing data. The AI models developed to automate this task usually classify the galaxies in classes of 3 or 10, and require massive hardware requirements due to the heavy computations involved. The used data was acquired from SDSS [1] and Galaxy Zoo [2], to make five Le-Net [3] based classification models with increasing computational requirements proportional to prediction accuracy. Contrary to what has often been assumed, classification models need not to be computationally demanding. Galaxy classification, to an accuracy of 82.7% can be achieved by a simple model. Although, for higher accuracy and classification of massive datasets, the hardware requirements and computational power have to be upgraded.**

I. Introduction

Edwin Hubble introduced his system of morphological (based on appearance) classification of galaxies in 1926. Morphological classification of galaxies plays an important role in the study of large scale structure of the universe. It is also used as a first hand tool to study the astrophysics of galaxies. Ever since the introduction of Hubble's system of classification, galaxies have been classified by manual inspection of their images. Manual classification was not a problem back then as only few galaxies were known.

Since the invention of CCDs (Charge Coupled Device) and with the advancement in computers, astronomy has seen a huge boom in collection of quality data. Especially in the last two decades, astronomy has become highly computational. When the Hubble Space Telescope captured the Hubble Ultra-Deep Field [4] image, which contained approximately ten thousand galaxies in a tiny portion of the sky, the knowledge of number of galaxies in the universe increased drastically. Extensive sky surveys like the Sloan Digital Sky Survey (SDSS) have catalogued thousands of galaxies and classifying them manually is infeasible. This brings us to automating the process of classification of galaxies using artificial intelligence techniques. In this paper we have developed a set of models for morphological classification of galaxies such that their accuracy increases with required computational power. Hence the lower end models can be implemented even in personal computers giving a fair enough accuracy.

II. Contributions

As evident from the Introduction, with the onset of high end telescopes, the data available increased exponentially. The classification of the images required manual identification by human vision, due to which image classification and data processing as a whole became a cumbersome and tedious task. With the onset of AI, there were several models developed to classify the images captured by the ground and star based telescope into number of different classes and objects. These model usually perform the computation on HDF format files that are a hierarchical structure to store the images, or ran on a massive dataset, while providing a classification accuracy of 90%. These model with all their pros were computationally challenging for normal public or a simple undergraduate research student to to run. These require massive hardware requirements due to high end computational demands. Also the current models classify the galaxies into three of ten classes.

- To make things flexible, the project focused on developing models that could be run on computationally weaker machines but still providing data with high enough accuracy to be used.
- The models also needed to be flexible enough for the user to modulate them and apply to the dataset of their own desired format to classify the images.
- This project fulfills this gap, by presenting a set of models with increasing computational demands, and hence increasing accuracy of prediction.
- These models are provided with their data classification and reduction programs that can be used to process a wide variety of data formats and convert them into HDF format that provided faster performance on hardware constrained machines.
- The models can also be modified and optimised to perform an n-index classification of not just galaxies but also star galaxy classification and many related requisites. The final model is a deep neural network which should provide an accuracy of 97% if run given the sufficient computational power and hardware support.

III. Background

The Hubble Morphological Classification using machine learning model has always been a pinnacle of research in the Galactic Astronomy. In the early days of this research area, there were not many satisfactory results obtained.

Storrie-Lombardi et al. (1992) [5] made a feed-forward neural network to perform a five-class classification of galaxies. With a training set of 1700 images and a test set of 3517 images, they achieved an accuracy of 64%.

Owens, Griffiths & Ratnatunga (1996) [6] used the same dataset provided by Storrie-Lombardi et al. (1992) [5] performed Decision Tree classification. They achieved an accuracy of 64.6 % using 5-fold cross-validation.

With the advancement in Artificial Intelligence, there has been significant contributions made in the field of galaxy classification.

Bazell & Aha (2001) [7] used three different classification algorithms that includes a Naive Bayes classifier, a neural network trained with backpropagation, and a decision-tree induction algorithm with pruning. The decision-tree algorithm achieved an accuracy of 78.55 % in three class galaxy classification of 800 galaxies.

Madgwick (2003) [8] used optical spectra of galaxies to classify them morphologically. He used Artifical Neural Networks (ANN) and achieved best accuracy of about 70 % for Elliptical + Lenticular galaxies and 83 % for Spiral + Irregular galaxies.

Calleja & Olac (2004) [9] used a neural network model and a locally weighted regression method and implemented homogeneous ensembles of classifiers. The homogeneous ensemble of locally weighted regression method achieved an accuracy of 91 % and 95 % on galaxy classification of three and two galaxy types respectively.

Maribel, Luis, et al.(2013) [10] achieved an accuracy of about 91 % for Random Forest Classifiers and 79 % for the Naive base classifier.

Khalifa, Nour Eldeen M., et al.(2017) [11] used a deep convolutional neural network to perform classification of galaxies in three categories (Elliptical Spiral and Irregulars). They achieved a test accuracy of 97.272 %.

Building upon the ideas of Khalifa, Nour Eldeen M., et al.(2017) [11], we have used CNN architecture to classify the galaxies in three categories. The Model MI is based on this paper.

IV. Material and Methods

A. Study Area

Galaxies [12] are a huge collection of gas, dust, and billions of stars and their solar systems, all held together by gravity. They play a fundamental role in the study of the universe. Classification of galaxy is the first step towards the comprehensive study of galaxies. In 1926, Edwin Hubble developed a classification scheme to classify galaxies based on there Morphology. The Hubble Galaxy Classification Scheme is best understood from the Tuning Fork Diagram given below:

In this paper, we have performed a three-class classification of galaxies. The details of the same are given below:

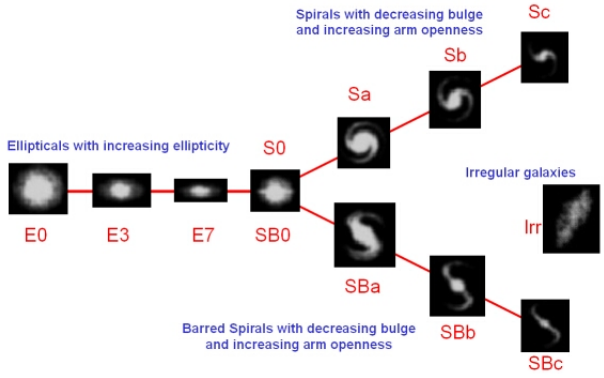


Fig. 1. The Hubble tuning fork diagram (Credits: astronomy.swin.edu.au)

1) Spiral Galaxies: They have a spiral structure that extend from center to the edge of the galaxy. They are further classified into barred *(Symbol: SB)* and unbarred spirals *(Symbol: S)*. Barred spirals have a bar like shape that extend from the center and the spiral arm begin from the other end of it.

Fig. 2. Unbarred Spiral galaxy: NGC 5194 (Credits: nasa.gov)

Fig. 3. Barred Spiral galaxy: NGC 1365 (Credits: eso.org)

2) Elliptical Galaxies: They have a nearly ellipsoidal shape and smooth image without any features. The naming convention is "E" followed by a positive integer. The greater the value of the integer, the more is the ellipticity of the galaxy. For example, E0 galaxy is more round (and less elliptical) than E7. There is another type of galaxy, called lenticular galaxy *(Symbol: S0)* which is an intermediate between ellipticals and spirals. These types of galaxy have a large-scale disc but no large-scale spiral arm. We are not concerned with this type of galaxy for this project.

Fig. 4. Elliptical galaxy: NGC 1365 (Credits: eso.org)

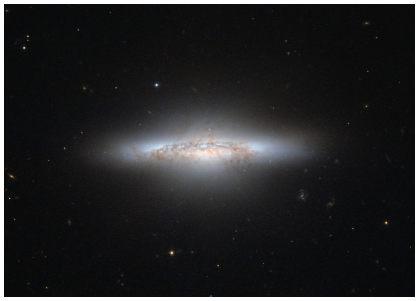


Fig. 5. Lenticular galaxy: NGC 5010 (Credits: nasa.gov)

3) Irregular Galaxies: They have no particular and distinct shape. They don't fall under any categories of Hubble classification scheme.

Fig. 6. Irregular galaxy: NGC 55 (Credits: eso.org)

B. Data

The data of classified galaxies was taken from the Galaxy10_DECals dataset [13]. The original Galaxy10 dataset was taken from Galaxy Zoo conataining images from SDSS. However, Galaxy Zoo later utilised images from DECam Legacy survey (DECals) which is a part of the DESI Legacy Imaging Surveys. The Galaxy10_DECals dataset is itself a combination of Galaxy Zoo Data Release 2 (GZ DR2) containing images from the DESI Legacy Imaging Surveys instead of SDSS and DECals campaign ab, c. The Galaxy10_DECals dataset contains about 441k unique galaxies out of which 17736 galaxies were selected for rigorous classification and the dataset which we have used contains these 17736 images.

1) Data reduction: The Galaxy10_DECals dataset contains 17736 colored images (taken in g, r and z bands) of the size 256×256 pixels and are sorted into 10 classes. The dataset that we used was stored in Hierarchical Dataset Format as Galaxy10_DECals.h5. The hierarchical data file contained arrays named 'images', 'ans', 'ra, 'dec', 'redshift' and 'pxscale'. Out of these only *images* and *ans* were useful for us. The *images* array contained the images while the *ans* array contained the information of labels (types) of galaxies. However, the *ans* array required some processing for the label information to be useful. Following are the classes into which the galaxies in the dataset are classified:.

TABLE I
TYPE DISTRIBUTION OF GALAXIES IN THE DATASET

Class	Type	No. of Galaxies
0	Disturbed Galaxies	1081
1	Merging galaxies	1853
2	Round Smooth Galaxies	2645
3	In-between Round Smooth Galaxies	2027
4	Cigar Shaped Smooth Galaxies	334
5	Barred Spiral Galaxies	2043
6	Unbarred Tight Spiral Galaxies	1829
7	Unbarred Loose Spiral Galaxies	2628
8	Edge-on Galaxies without Bulge	1423
9	Edge-on Galaxies with Bulge	1873

Since there were 10 classes of galaxies in the dataset and we needed to classify galaxies into only three classes - Irregulars, Ellipticals and Spirals, we merged the classes in the above table that fall into one of these three categories. Thus, classes 0 and 1 were merged to form the set of Irregular galaxies, classes 2-4 were merged to form the set of Ellipticals and the rest were merged to form the set of Sprials. With this, we obtained the type distribution of galaxies as shown in Figure 7.

On inspecting the galaxy images, we found that galaxies were centered in all images, but other unwanted objects were also present in the images. So we cropped the images to a size of 150×150 pixels to eliminate unwanted objects. This size was chosen keeping in mind computational constraints and also that galaxies shoudln't be cropped while removing other objects. Our models **Model 0** and **Model 0_1** were trained on this dataset. However, we found that both of them yielded less than expected accuracy. We strongly believed that the reason behind this would be the class of Merging galaxies. On inspecting the images of merging galaxies, we found (obviously) that each contained more than one galaxy of same or different type (spiral or elliptical). The presence of multiple galaxies in same image and their resemblance to spirals or ellipticals possibly confused our models in believing them to be irregulars while their features were similar to spirals or ellipticals. Also, many of the galaxies in the Disturbed galaxies class had resembled more with ellipticals or spirals than irregulars. In order to overcome the above two issues in the dataset, we completely removed the class of merging galaxies and manually classified the disturbed galaxies into irregulars, ellipticals and spirals. The distribution of galaxies

in this improved dataset is shown in Figure 15. Using this improved dataset, the accuracy of our models increased by about 10%.

The Galaxy Zoo open dataset on Kaggle [14] provide more than 1 lac unclassified galaxy images. We took around 40000 images from the galaxy zoo open dataset and intended to use our CNN model to classify these galaxy images. The size of these images were 424×424, while our model was trained on images of size 150×150. Hence, before classify images, we cropped them to the size of 150×150.

C. Methodology

An evolutional system of chronologically constrained set of models, each based on Le-Net architecture with increasing level of computational accuracy and proportionally increasing hardware and computational demand was prepared and trained for the classification of data in the following project.

The architecture of the overall project, codes and results can be divided into three major sections based on their operations.

1) Trained Models
2) Prospected Model MI
3) Prediction using the trained model

The trained models are the computationally easy and less hardware power demanding models, that we were able to develop and execute given the hardware constraints of the machines. Each model is named in a way that represents division and an evolution with respect to the previously developed and trained model. This section contains a total of 5 models named **Model 0, Model 0_1, Model 1, Model 1_1, Model 2**, with the Model 0, being the initial development and with the least accuracy and Model 2, being the most evolved, computationally demanding and most accurate of the *trained models.*

The Prospected Model MI is a deep convolutional network with several layers of convolution, pooling, flattening and filter. The existence of such complexity makes its execution virtually *impossible* on the machines available to us. Given the hardware constraints and hence the lack of computational power, the Model MI was not able to execute or train itself, hence remains to be a "prospected model". The accuracy of model MI is believed to be approximately 96% on any n-index classification of galaxies irrespective of the quality of the images but constrained by the number of images availed to the model for training.

The trained models were run on a dataset containing unclassified images from the defined dataset, to generate a label for each of the image based on the model's learned parameters. The images and the corresponding generated label was then collected and compiled into a *.csv* file.

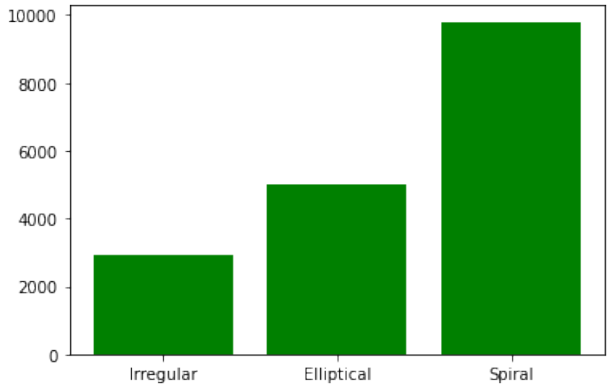
1) Model 0(Initial model without Convolution Layers): The dataset for this model was adapted directly from our acquired dataset [13]. The galactic images in the dataset are divided into three major groups, i.e.

1) Irregular
2) Elliptical
3) Spiral

This dataset distinguished between two types of Irregular galaxies and cited them differently. They were subdivided into Diffused Irregular galaxies, which are galaxies with no definite shape and look like irregular blobs of debris with smudged edges embedded with bright light sources through a telescope. On the other hand, the second ones are the merger galaxies which can be considered two different galaxies (usually the same type but can be different as well) that are merging into each other or being thrown apart after a merger event. This category features more than one kind of galaxy (although morphologically distorted) in a single image.

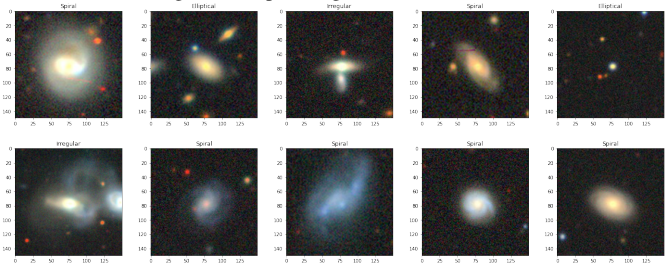
The data was imported into an HDF object and converted into a Pandas dataframe. The classification of the galaxies in the dataframe and the corresponding number of galaxies in each of the category can be retrieved from it. The graphical classification of the galactic images in the dataset can be observed as:

Fig. 7. Graphical Classification of galaxies in the dataset



The galactic classification and the graphical representation suggests that number of galaxies are the highest for the spiral type of galaxies the followed by elliptical galaxies and then finally the two types of Irregular galaxies (here in the same bar, as the dataset considers them the same). A sample of the galaxies, in the dataset can be seen below:

Fig. 8. Sample images from the dataset



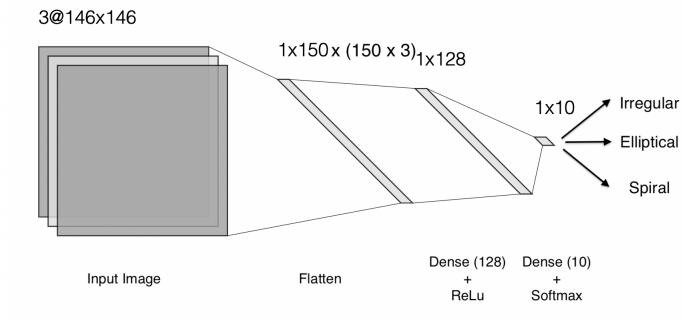
The above depicted images of the galaxies are some samples from the dataset, that was used to perform the model learning. The images are of the galaxies form approximately all the types as present in the universe, hence presents a better learning frame for the model.

The model architecture for Model 0, does not include a convolution layer as it was created with a mindset of classifying the galaxy images with minimal accuracy need and virtually no need of a high end computation machine with hardware arrangements.

The images are colorful hence are stored in the form of a three dimensional array with three, two dimensional arrays. The 2D array represents the number of pixels that define the width and the height of the galaxy images in the dataset. Since the images are colored, they are a combination of 3 matrices of RGB indices. Hence the images are a 3 dimensional matrix, being fed to the model. The first layer is the Flattening layer that converts this three dimensional array of the image into a simple one dimensional string of integers that contain the pixel values of the red, green and blue values of the image in sequential format. This flattened array is fed into a Dense layer, that performs a fully connected standard operation of the CNN architecture to yield a one dimensional array of length 128. A ReLu activation is applied to this layer for activation, and formation of the third layer. The third and final convolutional operation is the Dense operation consisting of 10 nodes on this 1×128 layer, to form a final one dimensional array of 10 integers. This final layer is passed through a *sigmoid* function that performs the sigmoid operation on it and classifies the final image into one of the 3 classes described. The model uses an "Adam" optimizer provided by the Keras system of the Tensorflow frame, with a learning rate of 0.001. The model zero monitored the loss by the 'sparse_categorical_crossentropy' method over the 'accuracy' matrix.

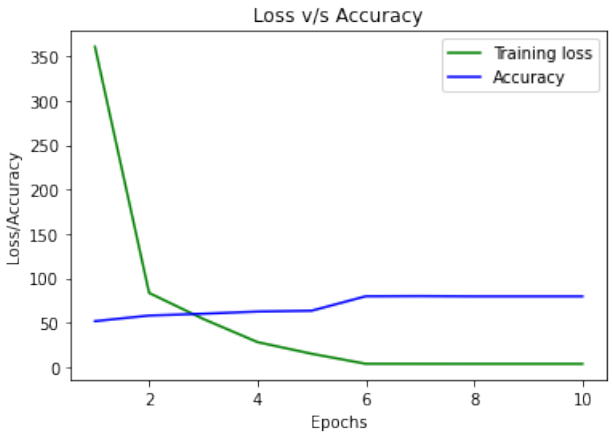
To optimize the learning rate according to each passing epoch, we utilised the 'ReduceLROnPlateau' module of the Keras system, which monitored the accuracy delivered by each of the passing epoch. If the accuracy did not show an increase of 0.01 within the lapse of an epoch the learning rate was reduced by a factor of hundredth automatically (via the 'auto' mode) to account for better learning. The architecture of the Model 0 can be visualised below,

Fig. 9. Architecture of Model 0



The model was run on the dataset for a total of 10 epochs in 500 batches. The accuracy increased with each epoch along with the gradual decrease in the learning loss which can be graphically visualised with the following plot.

Fig. 10. Loss v/s Accuracy plot for Model 0



This graphical representation demonstrates the learning curve of the model, and the subsequent decrease in loss function.

Post training, the model achieved an accuracy of approx 65%. Despite the absence of a convolutional layer, a simple flattening layer and two dense layers achieve an accuracy above 60%. The total number of trainable parameters involved in this model were about 8,641,418. The distribution and emergence of the number of parameters, is well classified and can be visualised from the model summary. The summary of the Model 0, is as

Fig. 11. Summary: Model 0

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
flatten (Flatten)            (None, 67500)             0

dense (Dense)                (None, 128)               8640128

dense_1 (Dense)              (None, 10)                1290

=================================================================
Total params: 8,641,418
Trainable params: 8,641,418
Non-trainable params: 0
_________________________________________________________________
```

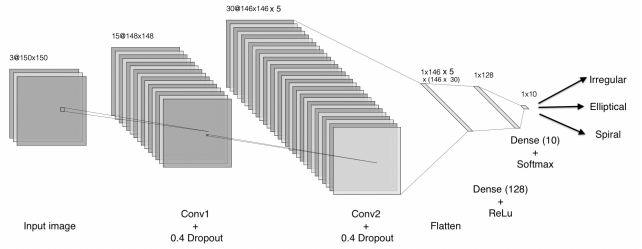
To further add on to the accuracy by utilising more of the hardware and hence computational, a upgradation of the Model 0 was created with two additional layers of convolution and named Model 0_1 (Model Class 0, upgrade 1).

2) Model 0_1(An improvement of Model 0 —— Addition of two Convolutional Layers): As the Model 0 was able to achieve a total accuracy of about 65% , a new model was created with an addition of two Convolution layers. The first layer constituted of a layer with total 5 filters to scan the galaxy image; The filters were in the form of 3×3 matrix that moved across the input layer with a stride of 2×2. The overall layers was activated by the trigonometric function 'tanh'.

This layer was followed by a dropout layer of index 0.4

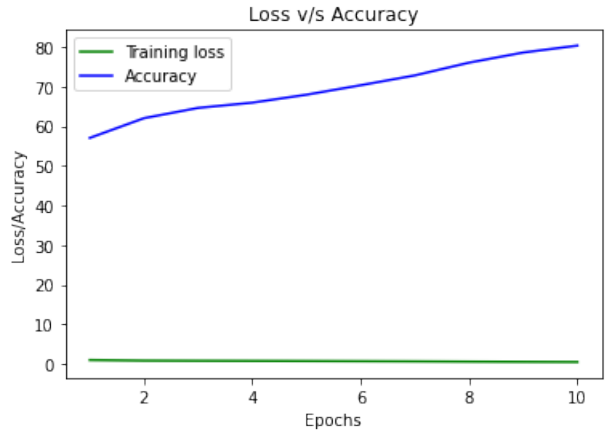
applied to the convoluted layer to avoid over-fitting of the model. The third layer was made of yet another convolution layer with total of 10 filters; The filters were in the form of 3×3 matrix that moved across the input layer with a stride of 2×2. The overall layers was activated by the trigonometric function 'tanh' and a dropout layer with an index of 0.4. This new addition of Convolution layers was achieved by the 'Conv2D' module of the Keras frame, and was followed by an architecture exactly similar to the one described in the Model_0. The visual representation of the model can be seen below

Fig. 12. Architecture of Model 0_1



The addition of two convolutional layers in the model added an accuracy of about 4%. The total accuracy of the Model 0_1 was observed to be approximately 68% when taught with a random seed of 10. There was a significant increase in the accuracy gained by the model with each of the epoch, and a subsequent steep drop in the learning loss of the model as graphically visualised in this plot,

Fig. 13. Accuracy v/s Loss plot for Model 0_1



The total number of trainable parameters involved in this model were about 1,660,898. The distribution and emergence of the number of parameters, is well classified and can be visualised from the model summary. The summary of the Model 0_1, is as

As seen above, the accuracy of the Model 0_1 is greater

Fig. 14. Summary: Model 0_1

```
Model: "sequential_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_2 (Conv2D)            (None, 74, 74, 5)         140
_________________________________________________________________
dropout_2 (Dropout)          (None, 74, 74, 5)         0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 36, 36, 10)        460
_________________________________________________________________
dropout_3 (Dropout)          (None, 36, 36, 10)        0
_________________________________________________________________
flatten_2 (Flatten)          (None, 12960)             0
_________________________________________________________________
dense_4 (Dense)              (None, 128)               1659008
_________________________________________________________________
dense_5 (Dense)              (None, 10)                1290
=================================================================
Total params: 1,660,898
Trainable params: 1,660,898
Non-trainable params: 0
_________________________________________________________________
```

than that of Model 0, by 4%. Despite the addition of two convolution layers, the accuracy could not be increased much.

The galaxy images were analysed manually and a assertion was made that the Merging galaxy types had two prominently visible and distinguishable galaxies in the image, and our model would detect the brighter or the larger galaxy and mark it as the label. Conclusively, the model was being trained as planned with respect to detecting the galaxies, but the existence of multiple galaxies in one image confused the model, and the 'apparent' accuracy was reduced. The word apparent states that the model was being trained good, and the decrease in the accuracy was merely due to multiple targets in the image. To overcome this, we decided to remove the merger galaxy class from the data set to see the absolute accuracy of the model training. More details about this can be found in the *Data* sub-section of *Material and Method* section.

3) Model 1(A model based on the improved and reduced dataset —— Without Convolution Layers): As discussed in Model 0_1, to remove the merged galaxy sets from the model, we simply removed it's label from the HDF file and skipped reading the images by providing proper indexing in the reading function. Since the images of the merger galaxies lied in the center of the data set, we imported the images before the mergers in one array and the images after it into a second array, and concatenated the two to form the input layer of the CNN model. With the removal of the merger galaxies, the number of Irregular galaxies was reduced. This step was supposed to decrease the confusion of the model and hence an attempt to increase the 'apparent' accuracy with an 'absolute' increase. The classification of galaxies and the distribution of the dataset can be visualised below with the following graphical representation.

Some sample images from the new dataset can be seen below.

As seen in the graph, the number of Irregular galaxies has dropped due to the elimination of the merger type of galaxies. To test the impact of the removal of the merger galaxies and

Fig. 15. Graphical Classification of galaxies in the renewed dataset

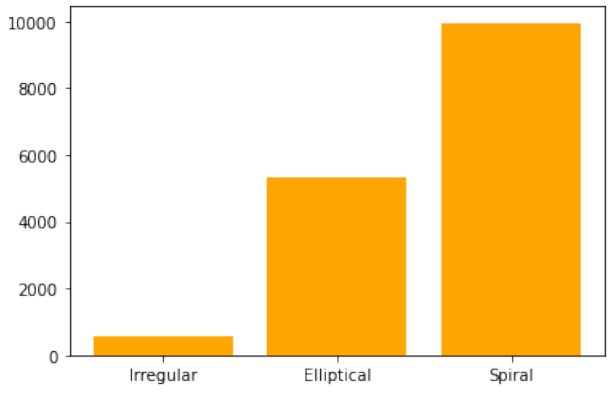
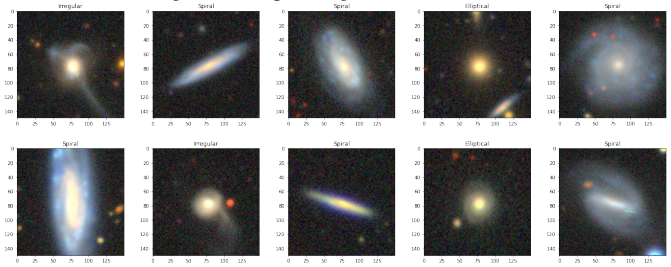
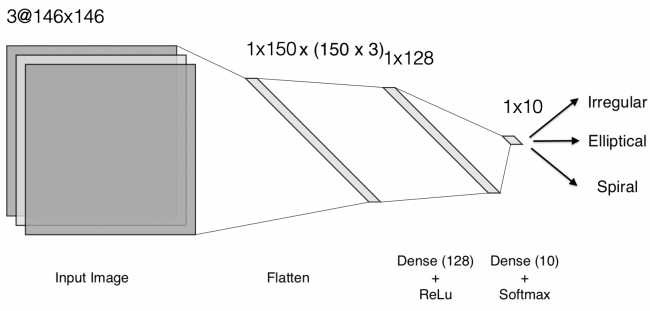


Fig. 16. Sample images from the dataset



the improvement of the dataset, a model exactly similar to Model 0, named Model 1 (Different Model Class as working on a new dataset) was tested on the new data. The architecture being exactly the same can be visualised below,

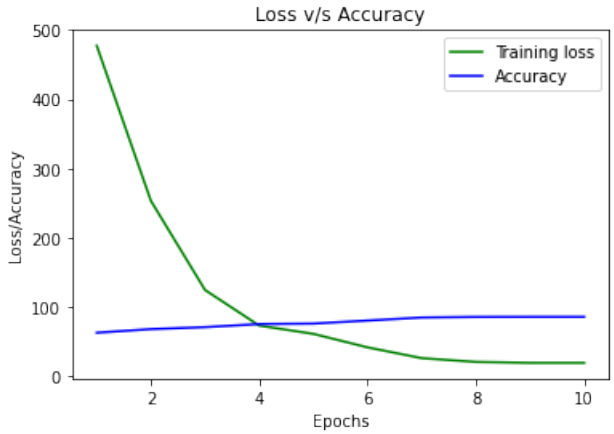
Fig. 17. Architecture of Model 1



The model yielded an accuracy score of approximately 77%. This was a massive increase in the amount of accuracy which proved our assertion; That the existence of merger galaxies in the dataset were causing confusion to the well trained model as they boasted more than one kind of galaxy in the image; true. The accuracy of the model gained high values and the learning loss fell following a steeper slope for this model. The accuracy and learning loss plot for the model can be graphically visualised by the following plot

The total number of trainable parameters involved in this model were about 8,641,418. The distribution and emergence

Fig. 18. Accuracy v/s Loss plot for Model 1



of the number of parameters, is well classified and can be visualised from the model summary. The summary of the Model 1, is as

Fig. 19. Summary: Model 1

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 flatten (Flatten)            (None, 67500)             0

 dense (Dense)                (None, 128)               8640128

 dense_1 (Dense)              (None, 10)                1290

=================================================================
Total params: 8,641,418
Trainable params: 8,641,418
Non-trainable params: 0
_________________________________________________________________
```

Now to test our assumption further, we utilised the Model 0_1 on the rectified dataset and found out the accuracy of the new model named Model 1_1 (As the model was trained on the same dataset but with different parameters and layers).

4) Model 1_1(A model based on the improved and reduced dataset —— With Convolution Layers): As seen in Model 1, the reduction and sub-classification of the dataset improves the accuracy of the model by reducing the training confusion. To increase the accuracy of the model, we apply a model exactly similar to the Model 0_1 on the new dataset and observe it's accuracy.

The architecture of this model can be visualised by this graphical representation

When this is trained over the new dataset, the model yields an accuracy score of approximately 80%. Therefore our assumption that the presence of merger type of galaxies was negatively impacting the accuracy of the model.

The accuracy and loss plot for the model can be visualised by the following plot

The total number of trainable parameters involved in this model were about 1,660,898. The distribution and emergence of the number of parameters, is well classified and can be

Fig. 20. Architecture of Model 1_1

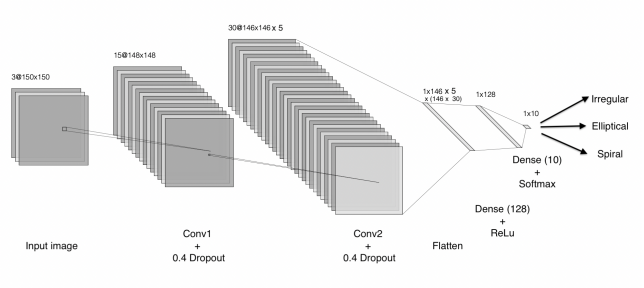
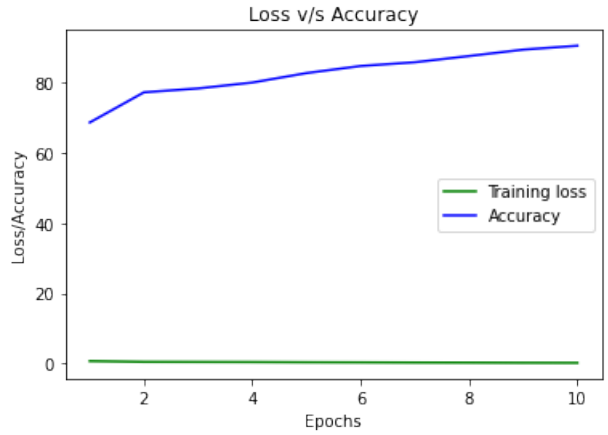


Fig. 21. Accuracy v/s Loss plot for Model 1_1



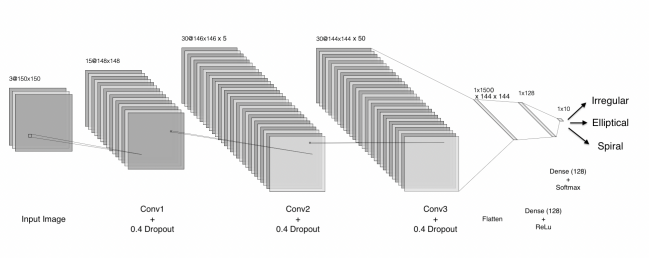
visualised from the model summary. The summary of the Model 1, is as

Fig. 22. Summary: Model 1_1

```
Model: "sequential_4"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_4 (Conv2D)            (None, 74, 74, 5)         140
_________________________________________________________________
dropout_4 (Dropout)          (None, 74, 74, 5)         0
_________________________________________________________________
conv2d_5 (Conv2D)            (None, 36, 36, 10)        460
_________________________________________________________________
dropout_5 (Dropout)          (None, 36, 36, 10)        0
_________________________________________________________________
flatten_4 (Flatten)          (None, 12960)             0
_________________________________________________________________
dense_8 (Dense)              (None, 128)               1659008
_________________________________________________________________
dense_9 (Dense)              (None, 10)                1290
=================================================================
Total params: 1,660,898
Trainable params: 1,660,898
```

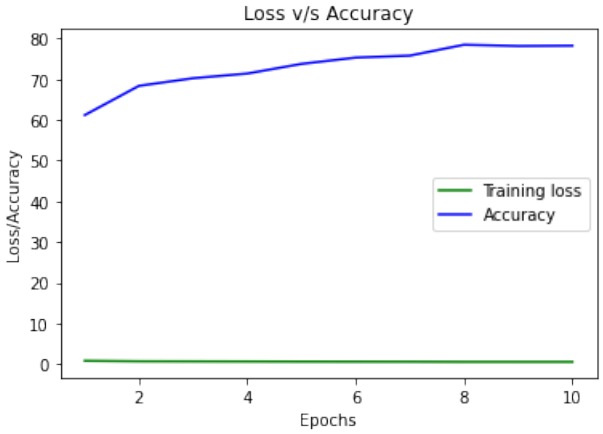
5) Model 2(An improvement of Model 1_1 —— Adding more Convolution Layers): To further increase the computational power and hence the accuracy of the model, we included an additional layer of convolution. This layer featured a ten filter Convolutional function layer, with kernel size of 3×3. The filters here move with a stride of 2×2 over the image, activated by a 'tanh' function. The architectural structure of the Model 2, can be seen in the visual representation below,

Fig. 23. Architecture of Model 2



When this is trained over the new dataset, this model yields an accuracy score of approximately 83%, which is the best score achieved by the models yet trained. We tried to add more layers of convolution and dropout to further increase the accuracy score, but were unable to do so due to hardware and computational power constraint.
The accuracy and loss plot for the model can be visualised by the following plot

Fig. 24. Accuracy v/s Loss plot for Model 2



The total number of trainable parameters involved in this model were about 1,660,898. The distribution and emergence of the number of parameters, is well classified and can be visualised from the model summary. The summary of the Model 1, is as

These were the five trained models created during the project, the final model i.e. Model MI (abbreviation of Model Mission Impossible) is a deep network containing approximately 50 times more parameters than Model 2, and hence requiring a high computational platform and hardware requirements.

6) Model MI—— Model Mission Impossible: This model is a deep neural network that is supposed to be trained on the cleaned or the uncleaned data set that we created, to

Fig. 25. Summary: Model 1_1

```
Model: "sequential_4"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d_4 (Conv2D)           (None, 74, 74, 5)         140

 dropout_4 (Dropout)         (None, 74, 74, 5)         0

 conv2d_5 (Conv2D)           (None, 36, 36, 10)        460

 dropout_5 (Dropout)         (None, 36, 36, 10)        0

 flatten_4 (Flatten)         (None, 12960)             0

 dense_8 (Dense)             (None, 128)               1659008

 dense_9 (Dense)             (None, 10)                1290

=================================================================
Total params: 1,660,898
Trainable params: 1,660,898
```

gain an accuracy of approximately 96%. The model should be powerful enough to not confused by the merger kind of galaxies and hence be able to provide high accuracy on the unfiltered dataset as well.
The model consists of the following layers

- A convolutional layer with 256 filters, each of the size of a 3×3 matrix, scans the image with the stride of 1. The input layer is the standard 3D image matrix from our dataset.
- A dropout layer with index 0.4, added to decrease the model complexity and avoid over-fitting.
- A second convolutional layer with 256 filters, each of the size of a 3×3 matrix, scans the image with the stride of 1. The input layer is the the dropped feature map created by the first convolutional layer.
- A Batch Normalization layer is added to make artificial neural networks faster and more stable through normalization of the layers' inputs by re-centering and re-scaling.
- A ReLu activation layer is added to add linearity to the data layer.
- A max pooling layer is added to the model, that moves on the input layer in the form of a 2×2 matrix and extracts the maximum of the four values it scans on the input layer.
- A third convolutional layer with 256 filters, each of the size of a 3×3 matrix, scans the image with the stride of 1. The input layer is the the Max Pooled layer created by max pooling of the activated batch normalization layer.
- A dropout layer with an index 0.25 added to avoid over-fitting in the model.
- A fourth convolutional layer with 128 filters, each of the size of a 3×3 matrix, scans the image with the stride of 1. The input layer is the the dropped max pooling layer.
- The three step process of Batch Normalization, ReLu activation and Max-pooling is repeated on the layer.
- A fifth convolutional layer with 128 filters, each of the size of a 3×3 matrix, scans the image with the stride of 1, followed by a 0.25 indexed dropout layer.
- The sixth and the last convolutional layer is added with 128 filters, the kernel size of 3×3 and a stride of 1.
- This is followed by Batch-Normalization, ReLu activation and a 'Global Max-Pooling of the model and a 0.25 indexed dropout layer.
- The final layer is flattened and a dense operation with 128 nodes is performed on it, twice with ReLu activation and subsequent addition of dropout layers.
- A final dense layer of 37 nodes reduces the overall structure to be fed into a sigmoid function that performs the final classification task.

The model being computationally demanding was not able to run due to hardware constraints. Although, if given sufficient power and input it should be able to classify the galaxy images with an accuracy of approximately 96-97%.

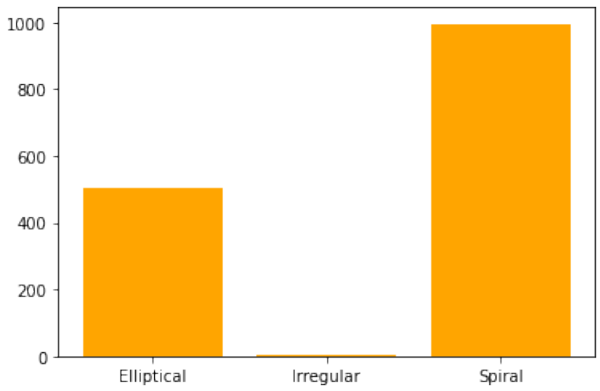
V. Results

We made five CNN models that can classify galaxies in three different categories (Spiral, Elliptical and Irregulars). Model 2 achieved the best accuracy of 82.7%, when tested on a galaxy dataset of 3971 images.
We classified 1500 unclassified galaxy images from Galaxy Zoo dataset available on Kaggle. The results of predicted types were saved in a CSV file called "Predicted_types.csv" with the corresponding GalaxyID. The resulted percentage composition of galaxies is found to be:

1) Spiral: 66.4 %
2) Elliptical: 33.4 %
3) Irregular: 0.2 %

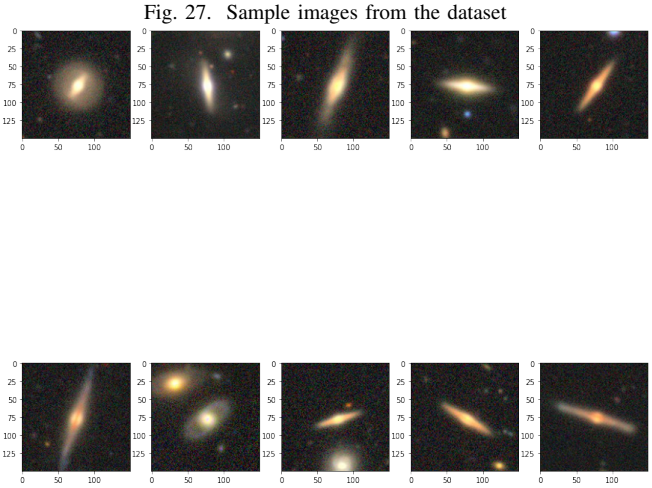
The above classification and distribution of the galaxies as predicted by our model can be visualised in the graphical representation below,

Fig. 26. Summary: Model 1_1



Some of the sample images classified by the trained model are represented below

Fig. 27. Sample images from the dataset



VI. Discussion

One of the main features of our project is that we could build a model that was less computationally demanding, could be run on a personal computer and yet achieve a decent accuracy. The accuracy of our best model (82.7%) is more than many of the old models that were developed by researchers. Potentially, we could have achieved a better accuracy if we had more good quality labelled data.

The abundance of different types of galaxies in the universe that we found above is in agreement with the theoretical estimates of the same. The table below summarizes the observed and theoretical percent abundance of spirals, ellipticals and irregulars in the universe.

Galaxy Type	Observed percentage	Theoretical percentage
Spiral	66.4	70-75
Elliptical	33.4	15-20
Irregulars	0.2	5

This result is very good when considering that our data had some bias and that the accuracy of our model was not the best that we could achieve. We also couldn't classify all 40000 unclassified images that we had due to computational constraints, else we would have got a better observed abundance of galaxies in the universe. This is the significance of our results.

VII. Conclusion

The CNN model was successful in classifying the previously unclassified galaxies and obtained the best accuracy of 82.7%. The estimated composition of galaxy types in the universe matched with the actual theoretical composition. The results couldn't be further improved because of unavailability of classified galaxy data. We were not able to run Model MI because of insufficient computational aids available to us. However, we strongly believe that if we were provided with a machine compatible enough to run Model MI, we could have achieved an accuracy of more than 95%. The model can be further improved to classify galaxies into more classes. We aspire to upgrade the model from a 3-class classification model to an n-class classification model.

References

[1] Sdss.org. 2021. SDSS. [online] Available at: ¡https://www.sdss.org/¿ [Accessed 22 November 2021].

[2] Zooniverse.org. 2021. Zooniverse. [online] Available at: ¡https://www.zooniverse.org/projects/zookeeper/galaxy-zoo¿ [Accessed 22 November 2021].

[3] Yann.lecun.com. 2021. MNIST Demos on Yann LeCun's website. [online] Available at: ¡http://yann.lecun.com/exdb/lenet/¿ [Accessed 22 November 2021].

[4] Beckwith, S., Stiavelli, M., Koekemoer, A., Caldwell, J., Ferguson, H., Hook, R., Lucas, R., Bergeron, L., Corbin, M., Jogee, S., Panagia, N., Robberto, M., Royle, P., Somerville, R. and Sosey, M., 2006. The Hubble Ultra Deep Field. The Astronomical Journal, 132(5), pp.1729-1755.

[5] Storrie-Lombardi, M., Lahav, O., Sodre, L. and Storrie-Lombardi, L., 1992. Morphological Classification of galaxies by Artificial Neural Networks. Monthly Notices of the Royal Astronomical Society, 259(1), pp.8P-12P.

[6] Owens, E., Griffiths, R. and Ratnatunga, K., 1996. Using oblique decision trees for the morphological classification of galaxies. Monthly Notices of the Royal Astronomical Society, 281(1), pp.153-157.

[7] Bazell, D. and Aha, D., 2001. Ensembles of Classifiers for Morphological Galaxy Classification. The Astrophysical Journal, 548(1), pp.219-223.

[8] Madgwick, D., 2003. Correlating galaxy morphologies and spectra in the 2dF Galaxy Redshift Survey. Monthly Notices of the Royal Astronomical Society, 338(1), pp.197-207.

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

[10] M. Marin, L. E. Sucar, J. A. Gonzalez, and R. Diaz, A Hierarchical Model for Morphological Galaxy Classification, in Proceedings of the TwentySixth International Florida Artificial Intelligence Research Society Conference, 2013, pp. 438443.

[11] Khalifa, Nour Eldeen M and Taha, Mohamed Hamed N and Hassanien, Aboul Ella and Selim, IM, 2017. Deep galaxy: Classification of galaxies based on deep convolutional neural networks.

[12] www.spacetelescope.org. 2021. The Hubble tuning fork - classification of galaxies. [online] Available at: ¡https://esahubble.org/images/heic9902o/¿ [Accessed 22 November 2021].

[13] Astronn.readthedocs.io. 2021. Galaxy10 DECals Dataset — astroNN 1.1.dev0 documentation. [online] Available at: ¡https://astronn.readthedocs.io/en/latest/galaxy10.html¿ [Accessed 22 November 2021].

[14] Kaggle.com. 2021. Galaxy Zoo - The Galaxy Challenge — Kaggle. [online] Available at: ¡https://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge/¿ [Accessed 22 November 2021].