Morphological Classification of Galaxies using AI

Mehul Paithane
Roll. No: 18142
Department of EECS
Email: mehul18@iiserb.ac.in

Yuvraj Sharma
Roll. No: 19288
Department of Physics
Email: yuvraj19@iiserb.ac.in

Yatharth Pal
Roll. No: 19340
Department of Physics
Email: yatharth19@iiserb.ac.in

Abstract—**Morphological [1] Classification of galaxies helps us understand their physical composition, various properties and evolutionary traits of the Galaxy, Galaxy clusters [2] and Halos [3]. We have used a Convolutional Neural Network with a LeNet5[1] based architecture to classify these galaxies into 4 major groups based on their Morphologies: *Elliptical, Spiral, Edge-on spiral and Irregular*. We used the data-set from the Galaxy10[2] to train and test our models, and then we used the Galaxy Zoo from Kaggle[3] to check predictions from our iModel we have thus created. We updated the LeNeT5 [1] architecture and shaped it to fit our dataset. We achieved a maximum accuracy of 89.83% during the training and an accuracy of 70.66% during the testing of our classified data. Our model successfully predicted unclassified data as well.**

I. Introduction

Understanding how and why we are here is one of the fundamental questions for the human race. Part of the answer to this question lies in the origins of galaxies, such as our own Milky Way. Yet questions remain about how the Milky Way (or any of the other 100 billion galaxies in our Universe) was formed and has evolved. Understanding the different morphology of these vast galaxies give us an insight to their nature.Galaxies come in all shapes, sizes and colors: from beautiful spirals to huge elliptical. Understanding the distribution, location and types of galaxies as a function of shape, size, and color are critical pieces for solving this puzzle. Galaxies are huge gravity bound bodies consisting of all kinds of space matter ranging from space dust to small asteroids to large nebulaes, they contain the information of how this universe was born.

Edward Hubble in 1936 was responsible for developing a classification scheme which would divide the galaxies into various categories based on their morphologies and later expanded by Gérard de Vaucouleurs and Allan Sandage. This system was eventually called as the Hubble sequence. As technology has advanced and the sheer volume of data produced by telescopes have increased. many telescopes both on-ground and off-ground have been generating tons of data. In July 2007, astronomers from Oxford University had in their possession a data set of 1 million galaxies imaged by the Sloan Digital Sky Survey. But the galaxies in this data set needed to have their morphologies (shapes) classified in order to be used to better understand galactic processes. With so many galaxies, it would have taken an individual a thousand lifetimes to classify all of them. A large number of galaxies have remained unclassified, creating a bottleneck in astronomical research. Moreover, some of the features depicted are virtually impossible to detect as the image quality and angle at which the pictures can be taken are limited.

Galaxy Zoo project recruits a number of volunteers to describe the morphology of galaxies by analyzing their images. More specifically, the volunteers step-by-step answer several questions about the galaxies. With exponential growth in the data-sets available for deep space imagery, manual classification of galaxies into categories continues to became less feasible manually. Visually classifying the apparent morphologies of galaxies through the manual inspection of images is a laborious, time

[1] Study of shape or form

[2] structure that consists of hundreds to thousands of galaxies that are bound together by gravity

[3] extended, roughly spherical component of a galaxy that extends beyond the main, visible cluster

consuming task. A more efficient and automated method is needed to classify the millions of images of different galaxies gathered by these telescopes for further studies. The aim of this project is to classify any galaxy based on the morphological by training a CNN model. Automated classification techniques, in particular those utilising neural networks, have the ability to revolutionise the speed at which samples can be individually classified. Neural networks have been utilised in this field for some time, but only recently has the rapidly growing field of deep learning seen widespread applications in astronomy. One of the key architectures behind the success of deep learning, particularly in applications involving image recognition. Convolutional neural networks (ConvNets) have recently been demonstrating their superior performances in numerous image classification challenges and datasets.

A few papers have been published on this topic, using various features and machine learning techniques. Our implementation looks to build off of these already done projects and learn from them. CNNs have been successfully utilised to detect quasars and gravitational lenses, study bulge/disk dominance, detect stellar bars and classify different radio morphologies. Understanding the distribution, location and types of galaxies as a function of shape, size, and color are critical pieces for solving this puzzle. This report summarises the efforts towards creating a neural model to classify these galaxies based on 4 different morphology: Elliptical, Spiral, Edge-on spiral and Irregular.

Fig. 1. Galaxy; Image Credit: ESA/Hubble & NASA

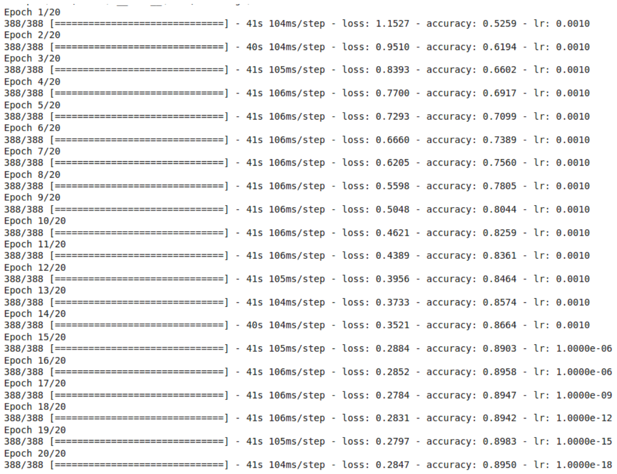
II. Contributions

This model can be used to automate and increase the efficiency of the process of classifying galaxies based on their morphology (Shape or structure). With the advent of large-scale structure surveys comes the need to process enormous amounts of data efficiently and accurately. This is crucial in the context of classifying galaxy morphologies, for which visual classification is intractable. Since a number of our parameters depend upon analyzing the ellipticity of the galaxy, the height and width of the galaxy, and the density of pixels in an area and these parameters assume that the galaxy is seen directly perpendicular, they are thrown off by these inclined galaxies. In order to solve this problem, the perpendicular view of these inclined galaxies would need to be projected from their images. This is a complicated procedure since it is difficult to tell if a galaxy is inclined, simply naturally very elliptic, or a barred spiral galaxy with only two short arms. Now, more than ever, galaxy morphology is a vibrant subject that continues to provide surprises as more galaxies are studied for their morphological characteristics across the electromagnetic spectrum. It is clear that a variety of effects are behind observed morphologies, including environmental density and merger/interaction history, internal perturbations, gas accretions, nuclear activity, secular evolution, as well as the diversity in star formation histories, and that a global perspective based on large numbers of galaxies will improve theoretical models and give a more reliable picture of galactic evolution. Galaxy morphology is a product of how galaxies formed, how they interacted with their environment, how they were influenced by internal perturbations, AGN, and dark matter, and of their varied star formation histories. This article reviews the phenomenology of galaxy morphology and classification with a view to delineating as many types as possible and how they relate to physical interpretations. The old classification systems are refined, and new types introduced, as the explosion in available morphological data has modified our views on the structure and evolution of galaxies. The main contributions that this project has made are as follows:

1) Morphology is still a logical starting point

for understanding galaxies. Sorting galaxies into their morphological categories is similar to sorting stars into spectral types, and can lead to important astrophysical insights. Any theory of galaxy formation and evolution will have to, at some point, account for the bewildering array of galactic forms.

2) Galaxy morphology is strongly correlated with galactic star formation history. Galaxies where star formation ceased giga-years ago tend to look very different from those where star formation continues at the present time. Classical morphology recognizes these differences in an ordered way.
3) Information on galaxy morphology, in the form of new types of galaxies, multi-wavelength views of previously known galaxy types, and higher resolution views of all or part of some galaxies, has exploded as modern instrumentation has super ceded the old photographic plates that were once used exclusively for galaxy classification.
4) Galaxy classification has gone beyond the realm of a few thousand galaxies to that of a million galaxies through the Galaxy Zoo project. Not only this, but Galaxy-Zoo has taken morphology from the exclusive practice of a few experts to the public at large, thus facilitating citizen science at its best. Galaxy Zoo images are also in color, thus allowing the recognition of special galaxy types and features based on stellar populations or gaseous emission. This project can be treated as a subset of galaxy-zoo as most of the information in this regard has already published in detail.
5) By identifying these galaxies we can find a suitable alternate galaxy which has a solar system similar to that of ours, also consisting of a planet similar to that of Earth, that can support life.
6) Finally, deep surveys with the Hubble Space Telescope have extended morphological studies well beyond the realm of the nearby galaxies that dominated early catalogues, allowing detailed morphology to be distinguished at unprecedented red-shifts.



```
Epoch 1/20
388/388 [==============================] - 41s 104ms/step - loss: 1.1527 - accuracy: 0.5259 - lr: 0.0010
Epoch 2/20
388/388 [==============================] - 40s 104ms/step - loss: 0.9510 - accuracy: 0.6194 - lr: 0.0010
Epoch 3/20
388/388 [==============================] - 41s 105ms/step - loss: 0.8393 - accuracy: 0.6602 - lr: 0.0010
Epoch 4/20
388/388 [==============================] - 41s 106ms/step - loss: 0.7700 - accuracy: 0.6917 - lr: 0.0010
Epoch 5/20
388/388 [==============================] - 41s 106ms/step - loss: 0.7293 - accuracy: 0.7099 - lr: 0.0010
Epoch 6/20
388/388 [==============================] - 41s 106ms/step - loss: 0.6660 - accuracy: 0.7389 - lr: 0.0010
Epoch 7/20
388/388 [==============================] - 41s 106ms/step - loss: 0.6205 - accuracy: 0.7560 - lr: 0.0010
Epoch 8/20
388/388 [==============================] - 41s 106ms/step - loss: 0.5598 - accuracy: 0.7805 - lr: 0.0010
Epoch 9/20
388/388 [==============================] - 41s 106ms/step - loss: 0.5048 - accuracy: 0.8044 - lr: 0.0010
Epoch 10/20
388/388 [==============================] - 41s 106ms/step - loss: 0.4621 - accuracy: 0.8259 - lr: 0.0010
Epoch 11/20
388/388 [==============================] - 41s 106ms/step - loss: 0.4389 - accuracy: 0.8361 - lr: 0.0010
Epoch 12/20
388/388 [==============================] - 41s 105ms/step - loss: 0.3956 - accuracy: 0.8464 - lr: 0.0010
Epoch 13/20
388/388 [==============================] - 41s 104ms/step - loss: 0.3733 - accuracy: 0.8574 - lr: 0.0010
Epoch 14/20
388/388 [==============================] - 40s 104ms/step - loss: 0.3521 - accuracy: 0.8664 - lr: 0.0010
Epoch 15/20
388/388 [==============================] - 41s 105ms/step - loss: 0.2884 - accuracy: 0.8903 - lr: 1.0000e-06
Epoch 16/20
388/388 [==============================] - 41s 106ms/step - loss: 0.2852 - accuracy: 0.8958 - lr: 1.0000e-06
Epoch 17/20
388/388 [==============================] - 41s 106ms/step - loss: 0.2784 - accuracy: 0.8947 - lr: 1.0000e-09
Epoch 18/20
388/388 [==============================] - 41s 106ms/step - loss: 0.2831 - accuracy: 0.8942 - lr: 1.0000e-12
Epoch 19/20
388/388 [==============================] - 41s 105ms/step - loss: 0.2797 - accuracy: 0.8983 - lr: 1.0000e-15
Epoch 20/20
388/388 [==============================] - 41s 104ms/step - loss: 0.2847 - accuracy: 0.8950 - lr: 1.0000e-18
```

Fig. 2. Model 1 Training

III. Background

IV. Materials and Methods

A. Dataset

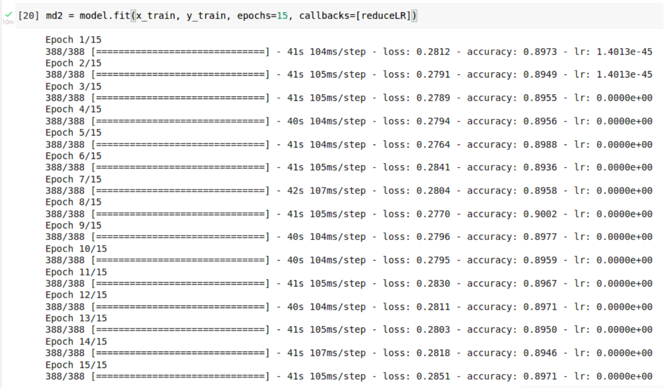
The original Galaxy10 dataset was created with Galaxy Zoo Data Release 2 where volunteers classify 270k of SDSS galaxy images where 22k of those images were selected in 10 broad classes using volunteer votes. GZ later utilized images from DESI Legacy Imaging Surveys (DECals) with much better resolution and image quality. Galaxy10 DECals has combined all three (GZ DR2 with DECals images instead of SDSS images and DECals campaign ab, c) results in 441k of unique galaxies covered by DECals where 18k of those images were selected in 10 broad classes using volunteer votes with more rigorous filtering. Galaxy10 DECals had its 10 broad classes tweaked a bit so that each class is more distinct from each other a with only 17 images in original Galaxy10 was abandoned.

B. Project

The main aim for our project is to classify Galaxies Morhpologically from their image data into 4 major groups

1) Spiral
2) Elliptical
3) Edge-on Spiral
4) Irregular

Using a CNN



```
[20] md2 = model.fit(x_train, y_train, epochs=15, callbacks=[reduceLR])

    Epoch 1/15
    388/388 [==============================] - 41s 104ms/step - loss: 0.2812 - accuracy: 0.8973 - lr: 1.4013e-45
    Epoch 2/15
    388/388 [==============================] - 41s 105ms/step - loss: 0.2791 - accuracy: 0.8949 - lr: 1.4013e-45
    Epoch 3/15
    388/388 [==============================] - 41s 105ms/step - loss: 0.2789 - accuracy: 0.8955 - lr: 0.0000e+00
    Epoch 4/15
    388/388 [==============================] - 40s 104ms/step - loss: 0.2794 - accuracy: 0.8956 - lr: 0.0000e+00
    Epoch 5/15
    388/388 [==============================] - 41s 104ms/step - loss: 0.2764 - accuracy: 0.8988 - lr: 0.0000e+00
    Epoch 6/15
    388/388 [==============================] - 41s 105ms/step - loss: 0.2841 - accuracy: 0.8936 - lr: 0.0000e+00
    Epoch 7/15
    388/388 [==============================] - 42s 107ms/step - loss: 0.2804 - accuracy: 0.8958 - lr: 0.0000e+00
    Epoch 8/15
    388/388 [==============================] - 41s 105ms/step - loss: 0.2770 - accuracy: 0.9002 - lr: 0.0000e+00
    Epoch 9/15
    388/388 [==============================] - 40s 104ms/step - loss: 0.2796 - accuracy: 0.8977 - lr: 0.0000e+00
    Epoch 10/15
    388/388 [==============================] - 40s 104ms/step - loss: 0.2795 - accuracy: 0.8959 - lr: 0.0000e+00
    Epoch 11/15
    388/388 [==============================] - 41s 105ms/step - loss: 0.2830 - accuracy: 0.8967 - lr: 0.0000e+00
    Epoch 12/15
    388/388 [==============================] - 40s 104ms/step - loss: 0.2811 - accuracy: 0.8971 - lr: 0.0000e+00
    Epoch 13/15
    388/388 [==============================] - 41s 105ms/step - loss: 0.2803 - accuracy: 0.8950 - lr: 0.0000e+00
    Epoch 14/15
    388/388 [==============================] - 41s 107ms/step - loss: 0.2818 - accuracy: 0.8946 - lr: 0.0000e+00
    Epoch 15/15
    388/388 [==============================] - 41s 105ms/step - loss: 0.2851 - accuracy: 0.8971 - lr: 0.0000e+00
```

Fig. 3. Model 2 Training

C. Importing Data

1) Mounting from Google Drive: Steps to follow while using drive from google.colab To mount your drive:

1) You will need to execute the code below
2) A link will be provided to get the authorization code
3) Copy the authorization code by signing in into the new page prompted
4) Enter the authorization code

The data we have is saved in h5 file format, here, we extract our data for further evaluation

We have retrieved the necessary data from our original dataset. i.e. images and ans

1) images : contains the RGB data for the image
2) ans : contain the class data, as to which class does the respective image belong to from the 10 classes specified

Now we have 17736 images available to us . As there is excess data in our image file of size 256x256, and the data required for our model training is centred for each image. So we will be cropping our images from size 256x256 to 144x144.

Here the labels are not specified in textual format. And for our project we are classifying the galaxies into 4 major groups

1) Spiral
2) Elliptical
3) Edge-on Spiral
4) Irregular

So we are using the variable labels to retrieve Galaxy class into a new list gal_class . After this we save our final dataset . Then we proceed to download our data i.e. Galaxy classes from ‘Gal_class.csv‘

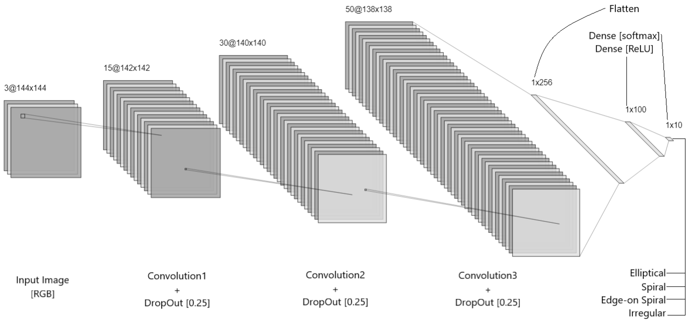
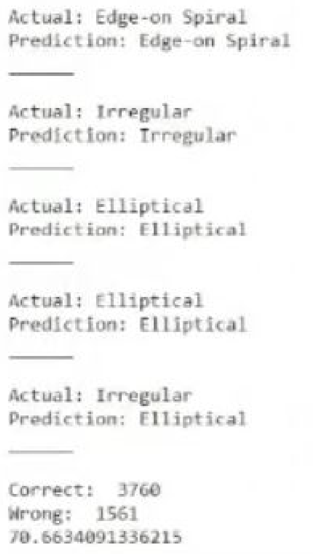


Fig. 4. CNN Architecture



```
Actual: Edge-on Spiral
Prediction: Edge-on Spiral
______
Actual: Irregular
Prediction: Irregular
______
Actual: Elliptical
Prediction: Elliptical
______
Actual: Elliptical
Prediction: Elliptical
______
Actual: Irregular
Prediction: Elliptical
______
Correct:  3760
Wrong:  1561
70.6634091336215
```

Fig. 5. Test Accuracy

V. Results

We ran our model and following is our result Accuracy and Loss v/s Epochs Plot.

Our model showed the following results during the testing of data form Kaggle Galaxy Zoo [3] data which was unclassified and pure for our model that is our model never saw the image previously. We had to crop this new data differently as this had a resolution of 424x424 resolution.

VI. Discussions

We used CNNs [a class of neural network] which helped us in extracting more features and marks out of the captured images. CNNs works on the

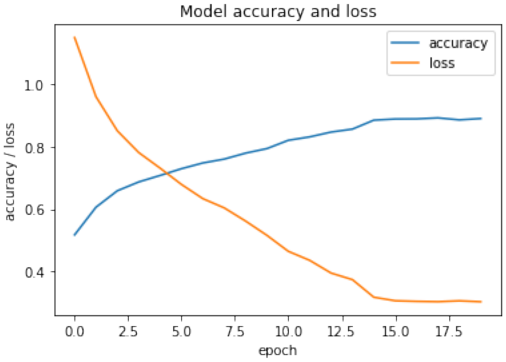


Fig. 6. Accuracy Loss Plot

principle of taking a dataset of images , training the model for a particular dataset and then helping us classify them for healthier segregation. In our project , we used iCNN to classify and help us segregate the types of galaxies by using a pre-loaded dataset of thousands of images of different types of galaxies . Not only in classification of galaxies but CNNs are also used in many applications like image recognition, face recognition, classification of different types of bacteria and microorganisms and video analysis . In order to prevent overfitting and the use of parameters that were not relevant for a specific classification type, we used the Dropout layer which nullifies the contribution of some neurons to the next layer and all others are left unmodified.

A. *Future Recommendations*

Future recommendations to enhance classification accuracy is by assimilating more training data and integrating versatile photometric features calculated in different spectra with morphological features. Executing bar-to-bulge ratio and the included curvature morphic features might also assist in boosting the accuracy on 4-category classification.

B. *Limitations*

A common problem faced in the algorithm was that the galaxies aren't always directly perpendicular to the line of sight of earth . This shows that galaxies in the dataset containing many images are at various inclinations. Thus circular spiral galaxies at an angle will look more like elliptical galaxies . And all our parameters considering the height , width , density of pixels and many more become useless in case of inclined galaxies . To get rid of this problem , we need a perpendicular view of all these inclined galaxies . But this further leads to more problems , as it will be more complex to find which galaxy is inclined and which one is a naturally elliptical or spiral galaxy . Another issue arose was that the size of galaxies in the images were not consistent and some images included foreign elements . Some galaxies took up the entire image without any dark space whereas some other images acquired very less space in image with expanse of space around them . Some galaxy images did not have any other celestial body except galaxy whereas some had many other celestial bodies which affect the training process of the model . Preprocessing that works to normalize the size of the galaxy within the image would serve to remedy this problem.

VII. Conclusion

A large amount ofi data is continuously getting produced by new generation telescopes which survey millions of galaxies covering all of the sky . LSST and JWT will provide data in millions which we will be unable to process by applying common methods . Some automated tools were developed specifically to deal with such a huge amount of information . CNN help us sort and classify millions of galaxies in a very brief period of time for which humans could take years and years . After the best models are trained and tested , a simple finetuning will be enough to revamp the new type of images for a reliable classification.

In this whole project , we gave a overall brief about galaxy morphological classification using Machine Learning methods, and also presented how the authors are trying to improve the accuracy reached by other works. These models will be applied in the new surveys for future studies in astronomy .

Acknowledgment

References

[1] LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W. and Jackel, L., 2021. Backpropagation Applied to Handwritten Zip Code Recognition.

[2] https://astro.utoronto.ca/ hleung/shared/Galaxy10/Galaxy10_DECals.h5

[3] https://www.kaggle.ciom/c/galaxy-zoo-the-galaxy-challenge/data

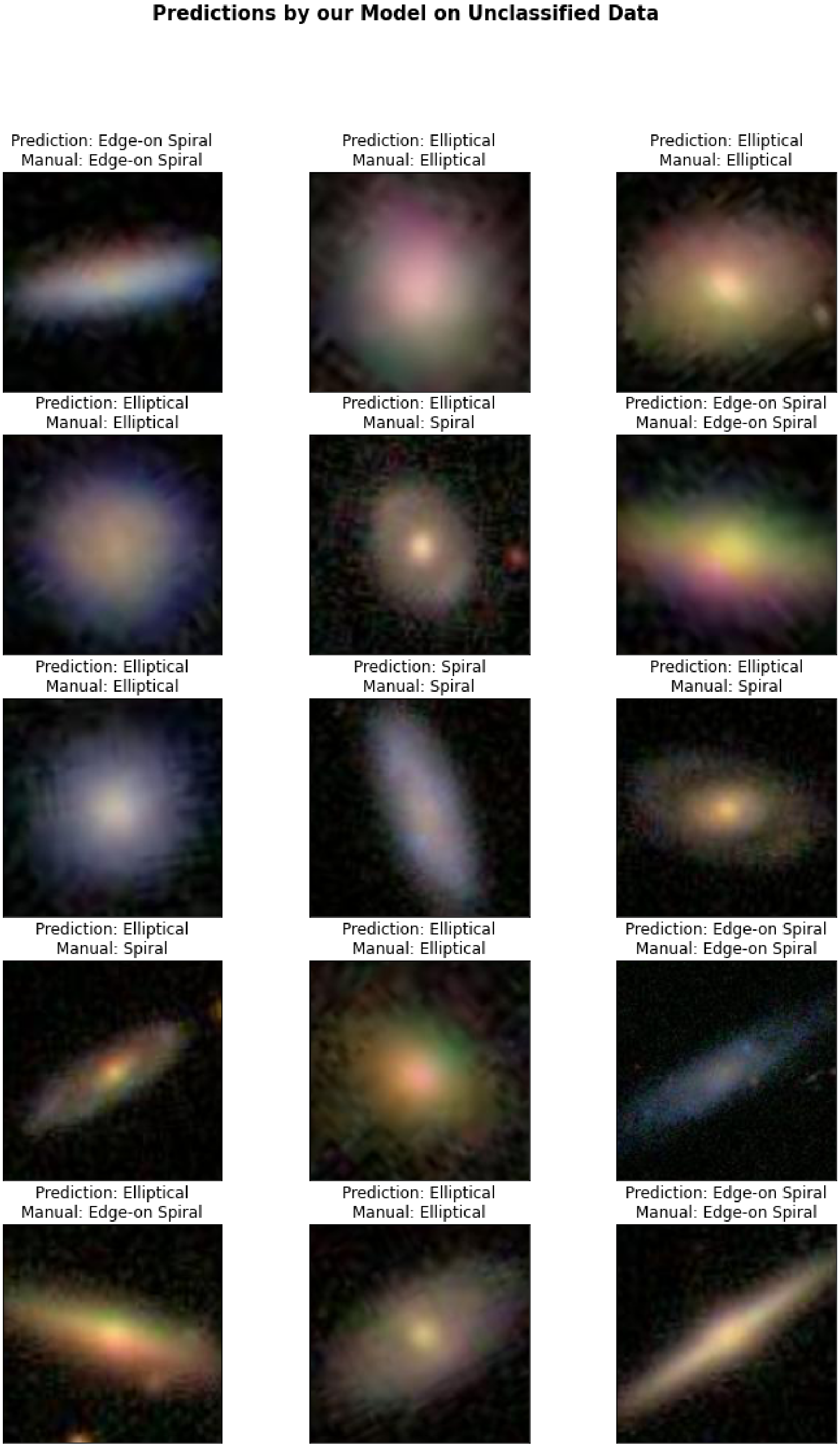


Fig. 7. Class Prediction