Predicting Galaxy Morphology: A Machine Learning Approach Leveraging Morphological Data to Refine Shape Classification and Insights into Galactic Evolution

Jayashree Kundu,
Research Scholar,
Dept. of Computer Science & Technology,
University of North Bengal
jayashree.cs.res.2023@gmail.com

Dr. Rakesh Kumar Mandal,
Associate Professor, , Dept. of Computer Science & Technology,
University of North Bengal
rakeshmandal@nbu.ac.in,

Dr. Tamal Sarkar, Scientific Research Officer-II, High Energy & Cosmic Ray Research Centre,
University of North Bengal,
ts.phys.edu2020@gmail.com,

Jayashree Kundu
Rakesh Kumar Mandal
Tamal Sarkar

Abstract:

The shape of a galaxy offers critical hints about its history, composition, and gravitational forces acting within and around it. This research introduces a novel approach to developing a computational code to predict the shapes of galaxies by leveraging Abell 1656(Coma-Cluster) data from SDSS and applying Random Forest as the machine learning algorithm to classify galaxies based on their photometric and positional properties. In the present work, we analyse the spectral data and its features that correlate with specific galactic structures to determine the shape of the Galaxies, count, and thus predict the shape of the galaxies for a new dataset. This paper helps classify galaxies into shape categories with high precision, even under challenging conditions like faint luminosity or overlapping structures by using Random Forest and supervised clustering. Our method advances traditional morphology classification by incorporating spectral data to refine shape predictions, providing a richer understanding of galactic evolution and dynamics. These insights enhance our knowledge of the interactions shaping galaxies over cosmic time, supporting future deep-field surveys and expanding the framework for studying the architecture of the universe.

Keywords: *Coma-Cluster, Galaxy, Morphology, Random Forest Model, Supervised Clustering.*

Abbreviations & Full Forms: *SDSS-Sloan Digital Sky Survey.*

1. Introduction:

The shape of a galaxy, also known as its morphology, provides vital information about its formation, evolution, and the physical processes going on within and around it. In 1926, Edwin Hubble introduced the Hubble Sequence, also known as the Hubble Tuning Fork Diagram, which

pioneered the classification of galaxies into morphological types (Dick, 2019)[1]. This sequence primarily focuses on spiral and elliptical galaxies, as these types represent the dominant structural forms in most observations. Spirals are characterized by their distinct gas-rich disk and arms, whereas ellipticals have a smooth, featureless shape consisting of older stars and lack significant gas or dust. Abell 1656(Coma-Cluster) is a famous Galaxy Cluster for its prominent and massive galaxies (over 1000 identified) located about 320 million light-years (~100 Mpc) away in the constellation of Coma Berenices (Mahajan, 2023)[2]. In this paper, we use the Sloan Digital Sky Server (SDSS)'s navigation tool (Gray, Szalay, Nieto-Santisteban, & Budavari, 2004)[3] to collect data (~300 data) on Abell 1656 which includes photometric imaging, spectroscopic measurements, and information across multiple filters (u, g, r, i, z). A color-color diagram is a graphical representation used in astronomy to study the properties of galaxies, or other celestial bodies based on their colors that are determined by measuring differences between different wavelength bands used by the Sloan Digital Sky Survey (SDSS), among which u-g and g-r color indices are significant in classifying galaxies. In the following section, we discuss computational code development with the help of Python Libraries and a robust machine-learning model, Random Forest (RF), which is very powerful in handling structured dataset.

2. Literature Review:

Python has become one of the most popular programming languages in astronomy due to its versatility, ease of use, and rich ecosystem of scientific libraries for astronomical research (Helfrich, 2019)[4] (Greenfield, 2011)[5]. Some prominent ones include NumPy for numerical computations, SciPy for scientific computing, matplotlib for data visualization, and Astropy for astronomical data analysis. Another important parameter for Python is that it is an open-source language that aligns well with the principles of openness and transparency in scientific research. This allows researchers to inspect, modify, and share the code freely (Morris, 2018)[6]. Python is an essential tool in computational astrophysics as it allows astrophysicists[3]to efficiently analyze

[1] Dick, S. J. (2019). The Galaxy Family. In *Classifying the Cosmos* (pp. 333-370). Springer, Cham
[2] Mahajan, S. (2023). *Exercises in Astronomical Data Analysis for Beginners.* Bhilai: OrangeBooks Publication
[7] Ivezić, Ž. C. (2020). Statistics, data mining, and machine learning in astronomy: a practical Python guide for the analysis of survey data (Vol. 8). Princeton University Press.
[8] Schmidt, W., & Volschow, M. (2021). Numerical Python in Astronomy and Astrophysics. Cham: Springer.
[9] Reza, M. (2021). Galaxy morphology classification using automated machine learning. Astronomy and Computing, 37, 100492.
[10]Coe, S. R. (2016). What Can I Observe in a Cluster of Galaxies?. In Deep Sky Observing. Springer, Cham, (pp. 115-132).
[11]Bahcall, N. A. (1999). Abell's Catalog of Clusters of Galaxies. Astrophysical Journal,, Vol. 525C,(Centennial Issue,), p. 873, 525, 873.
[12]Secker, J. (1996). Deep CCD photometry of the rich galaxy cluster Abell 1656: characteristics of the dwarf elliptical galaxy population in the cluster core. Publications of the Astronomical Society of the Pacific, 550.

data, model physical systems, and visualize results (Ivezić, 2020)[7]. Python frameworks like Scikit-learn, TensorFlow, and PyTorch enable the classification of galaxies, prediction of stellar properties, and analysis of complex datasets (Schmidt & Volschow, 2021)[8]. Various machine learning methods have already been applied to wide astrophysical phenomena. For example, Random Forest has been used to classify multiwavelength data (Reza, 2021)[9]. RF remains a keystone algorithm in machine learning due to its versatility, robustness, and ease of use. It is particularly suited for tasks requiring high accuracy, feature importance analysis, and resistance to overfitting. The Abell clusters were first cataloged by the American astronomer George O. Abell in 1958, examining the Palomar Sky Survey plates (Coe, 2016)[10] (Bahcall, 1999)[11]. Abell 1656 (Coma Berenices) is a dense, elliptical rich cluster with relatively scarce spiral galaxies (Secker, 1996)[12], especially in the central regions (Sreedhar, Rakos, Hensler, & Odell, 2009)[13]. From a large sample of Abell clusters with measured redshifts $z < 0.1$, a catalog of 48 probable superclusters is generated by employing a percolation technique to examine various properties of these superclusters in the universe (West, 1989)[14].

3. Research Gap:

Nowadays, a lot of applications of Artificial Intelligence (AI) supported languages and models are used for the analysis of big data. In this work, we have used Python which has many in-built packages like numpy, pandas, sklearn, matplotlib, astropy, etc. The package 'sklearn' has many AI features like 'Random Forest Classifier', 'accuracy_score' and many more. From the literature survey, we found that such tools have not been used earlier for this type of specific application.

4. Objectives:

The following are the objectives of this work: -

i) To calculate the type of the galaxy from the Coma cluster data,
ii) To count its type, whether it falls under the category of Spiral or elliptical,
iii) To predict the type of galaxy for a new upcoming dataset.

5. Methodology:

We have taken secondary data from the SDSS Server, which is an archive of the major multi-spectral imaging and spectroscopic redshift survey, which began in 2000 and is named after

[13]Sreedhar, Y. H., Rakos, K., Hensler, G., & Odell, A. P. (2009). Narrowband photometry and evolution of galaxies in Abell 1656 (Coma cluster) . Joint annual meeting of the Austrian physical society (p. 154). Austria: Swiss Physical Society (Switzerland).

[14]West, M. J. (1989). On the morphology of superclusters. Astrophysical Journal, Part 1 (ISSN 0004-637X), vol. 347, Dec. 15, 1989, p. 610-626., 347, 610-626.

Alfred P. Sloan. Astronomers use the digital data for further study and research. Our overall methodology used for our analysis is as follows.

(i) Preparing Table (i.e., Table 1) in the form of Comma Separated File for Abell 1656 galaxy cluster having ten columns, containing essential information like Serial number, Object-ID, RA, Dec, u, g, r, i, z filters, and redshifts.

(ii) Reading the data of Abell 1656 galaxy cluster.

(iii) Finding the color indices by calculating the difference between different filters.

(iv) Finding and counting the shapes of galaxies (using u, g, and r filters).

(v) Data Pre-processing (i.e., removing 'NAN' values) before creating the model for prediction.

(vi) Predicting through Random Forest Model.

(vii) Storing the results to a destination folder.

6. Data Collection:

The recent parameters/data of 322 galaxies of the Coma galaxy cluster have been taken from the SDSS by using its navigation tool, shown in Figure 1, through the following Data Collection process.

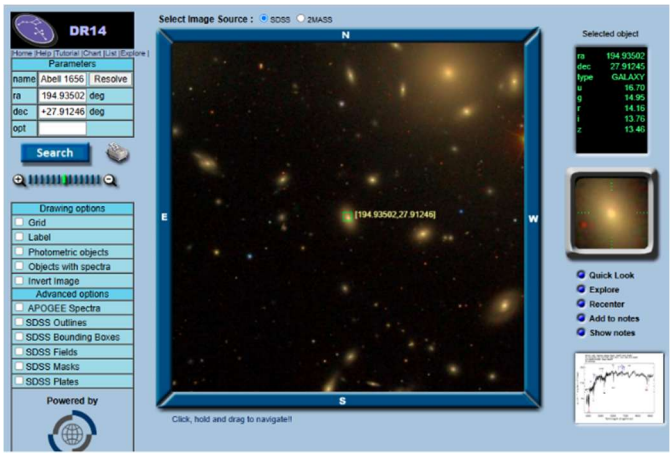


Figure 1: SDSS Navigation Tool

- Step 1: Go to 'web browser' and type 'https://skyserver.sdss.org/'
- Step 2: Choose 'Data Release 14' from a dropdown list.
- Step 3: Go to 'Data Access' and click on the 'Navigate' option.
- Step 4: Go to 'Parameter' and write the name of the cluster (here, in this case, it is 'Coma'/'Abell 1656')
- Step 5: Click on the 'Resolve' button to observe a high concentration of galaxies.
- Step 6: Select 'Objects with spectra'.
- Step 7: Select a galaxy from the SDSS window.

- Step 8: Click on 'Add to notes'.
- Step 9: Continue the process from steps 6 to 8 for other galaxies.
- Step 10: Click on 'Show notes' which will open 'Sky Server Notebook' as shown in Figure 2.
- Step 11: Finally, export the list as CSV and use the same for analysis, listed in Table 1.

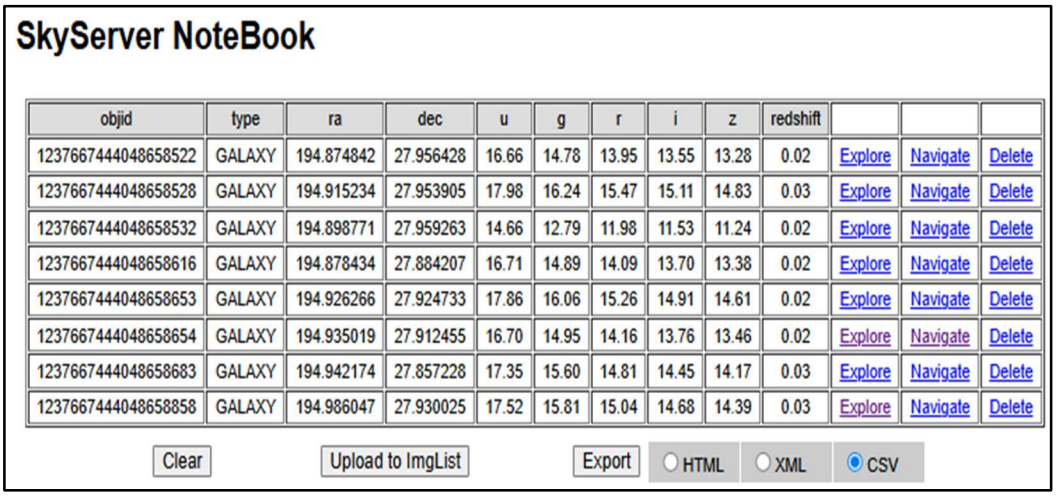


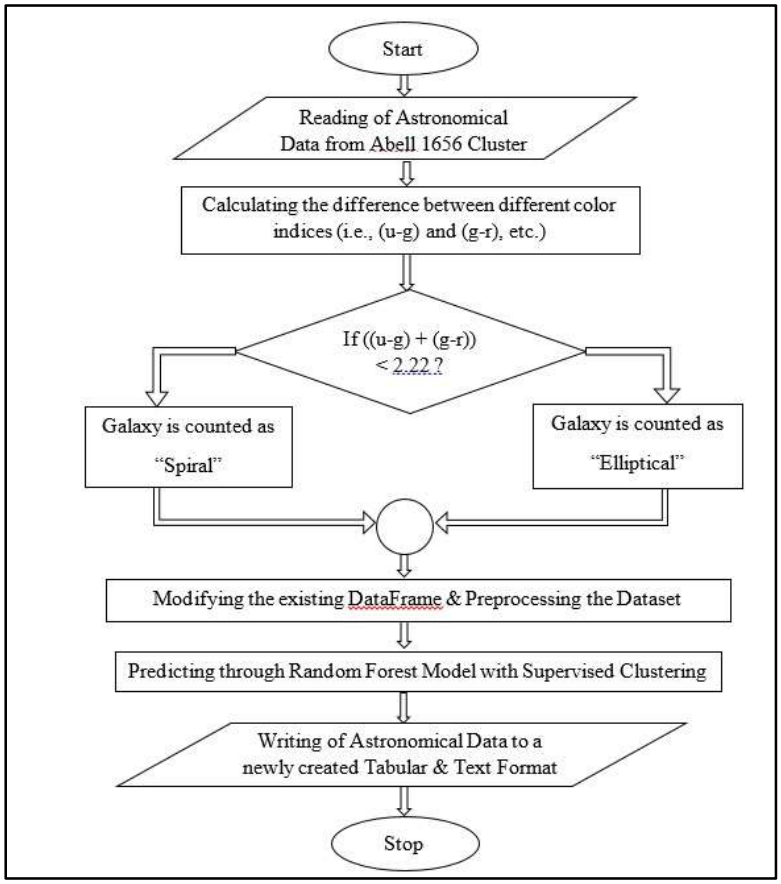
Figure 2: Synopsis of Sky Server Notebook

Table 1: Data(Abell 1656) captured from SDSS

SL. No.	Object-ID	RA	Dec	u	g	r	i	z	redshift
1	1237667324334506010	194.758627	28.115665	16.5102	15.55231	15.12759	14.83944	14.6472	0.03141
2	1237667444048593359	194.83920	27.97356	18.77836	17.07322	16.36368	16.02497	15.7741	0.020206
3	1237667444048658528	194.91523	27.95390	17.97682	16.24361	15.46985	15.10885	14.8285	0.0267883
…	……	…..	…..	…..	…..	….	….	….	……
321	1237667444048658531	194.93417	27.958413	18.29931	16.56185	15.82116	15.46881	15.20135	0.02303317
322	1237667444048658616	194.878434	27.884207	16.71277	14.89146	14.08934	13.69809	13.38376	0.01570999

7. Study Conducted:

Based on the methodology, we prepared the flow diagram, Flowchart 1, which takes the SDSS data of the Coma cluster, processes those data to classify using an optimum color separator, and depending upon that classification it can predict a new galaxy of that cluster.



Flowchart 1: Determination of the Shapes of Galaxy in Galaxy-Cluster & Predicting using Random Forest

8. Interpretation of Results:

In Table 2, [4]the summary of Findings, we have reported the essential results we got after implementing the codes using Python based on our research methodology. Further, in Figures 3 ,4 & 5 we have shown the classification of galaxies of the Abell 1656, Abell 2225, and Abell 1689 galaxy clusters using the optimum color separator (Strateva & et, al, 2001)[15]. From this classification (Tojeiro & et al., 2013)[16] we can visualize that only a smaller fraction of clusters are Spiral and the rest are predominantly older Elliptical galaxies.

[15]Strateva, I., & et, al. (2001). Color separation of galaxy types in the Sloan Digital Sky Survey imaging data. The Astronomical Journal, 122(4), 1861.
[16]Tojeiro, R., & et al. (2013). The different star formation histories of blue and red spiral and elliptical galaxies. Monthly Notices of the Royal Astronomical Society, 432(11), 359-373. doi:10.48550/arXiv.astro-ph/0107201

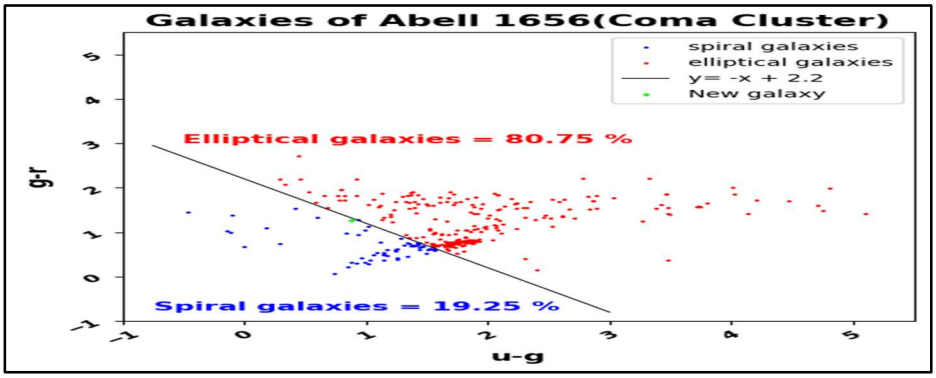


Figure 3: Galaxy Classification in Coma (Abell 1656) Cluster

As, the goal is to classify the structured dataset into distinct categories (e.g., elliptical or, spiral), Random Forest(RF), a machine-learning algorithm, has been used here, where the independent variables are chosen in such a way that it gives a better prediction (here, we are getting ~96-98% prediction score by having both the in-built Photometric features and some important derived features (e.g. color(u-g), color(g-r))). We can see a "lime" dot, a new galaxy, in those clusters(in Figure 3, it indicates a 'spiral', in Figure 4 it is an 'elliptical', and in Figure 5 it indicates also an 'elliptical') whose galaxy type is being predicted through the RF model. The following table (Table 2) depicts the summary of our findings.

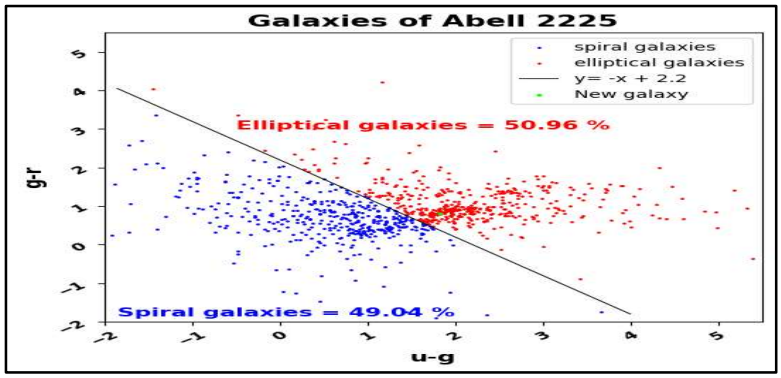


Figure 4: Galaxy Classification of Abell 2225 Galaxy Cluster

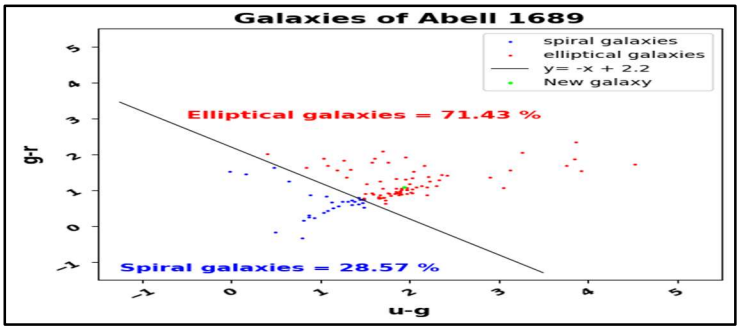


Figure 5: Galaxy Classification of Abell 1689 Galaxy Cluster

Table 2: Summary of findings

Name of Galaxy Cluster	Distance from Earth's base	No. of SDSS Data (captured)	No. of		Percentage(%) of		Prediction score(using RF)
			Spiral Galaxies	Elliptical Galaxies	Spiral	Elliptical	
Abell 1656	~320 million Ly	322	62	260	19.25	80.75	96.9%
Abell 2225	~2.4 billion Ly	889	436	453	49.04	50.96	97.75%
Abell 1689	~2.2 billion Ly	105	30	75	28.57	71.43	96.87%

9. Conclusion:

By training the algorithm/model using galaxies from one cluster (Abell1656) and applying the trained model (which has the same 'test_size', 'random_state', and 'n_estimators') to that galaxy cluster and other two(i.e., Abell 2225 and Abell 1689), we show how the algorithm can cleanly separate early-type and late-type galaxies with an accuracy range of 96-98%.

References:

Bahcall, N. A. (1999). Abell's Catalog of Clusters of Galaxies. *Astrophysical Journal,, Vol. 525C,*(Centennial Issue,), p. 873, 525, 873.

Coe, S. R. (2016). What Can I Observe in a Cluster of Galaxies?. In Deep Sky Observing. *Springer, Cham*, (pp. 115-132).

Dick, S. J. (2019). The Galaxy Family. In S. J. Dick, *In Classifying the Cosmos* (pp. 333-370). Springer, Cham. doi:https://doi.org/10.1007/978-3-030-10380-4_14

Gao, D. L. (2010). Multiwavelength study of nearly face-on low surface brightness disk galaxies. *Research in Astronomy and Astrophysics, 10(12)*, 1223. doi:DOI 10.1088/1674-4527/10/12/004

Gauthier, A. J. (2016). Galaxy morphology classification. *Lecture Notes, Stanford University*, 16.

Gray, J., Szalay, A., Nieto-Santisteban, M., & Budavari, T. (2004, 02). *SDSS DR8 Navigate Tool*. Retrieved from skyserver.sdss.org: https://skyserver.sdss.org/dr8/en/tools/chart/navi.asp

Greenfield, P. (2011, July). What python can do for astronomy. *In Astronomical Data Analysis Software and Systems XX, Vol. 442*, p. 425.

Helfrich, G. (2019). 4 Python tools for getting started with astronomy. *From opensource.com:https://opensource.com/article/19/10/python-astronomy-open-data*.

Ivezić, Ž. C. (2020). *Statistics, data mining, and machine learning in astronomy: a practical Python guide for the analysis of survey data* (Vol. 8). Princeton University Press.

Katgert, P. (2001). *Abell Clusters,In Encyclopedia of Astronomy & Astrophysics* (1st ed.). IOP Publishing. doi:https://doi.org/10.1201/9781003220435

Mahajan, S. (2023). *Exercises in Astronomical Data Analysis for Beginners.* Bhilai: OrangeBooks Publication.

Morris, B. M. (2018). Astroplan: an open source observation planning package in Python. *The Astronomical Journal, 155(3)*, 128. doi:DOI 10.3847/1538-3881/aaa47e

Reza, M. (2021). Galaxy morphology classification using automated machine learning. *Astronomy and Computing, 37*, 100492.

Schmidt, W., & Volschow, M. (2021). *Numerical Python in Astronomy and Astrophysics.* Cham: Springer.

Secker, J. (1996). Deep CCD photometry of the rich galaxy cluster Abell 1656: characteristics of the dwarf elliptical galaxy population in the cluster core. *Publications of the Astronomical Society of the Pacific*, 550.

Sreedhar, Y. H., Rakos, K., Hensler, G., & Odell, A. P. (2009). Narrowband photometry and evolution of galaxies in Abell 1656 (Coma cluster) . *Joint annual meeting of the Austrian physical society* (p. 154). Austria: Swiss Physical Society (Switzerland).

Strateva, I., & et, al. (2001). Color separation of galaxy types in the Sloan Digital Sky Survey imaging data. *The Astronomical Journal, 122*(4), 1861.

Tojeiro, R., & et al. (2013). The different star formation histories of blue and red spiral and elliptical galaxies. *Monthly Notices of the Royal Astronomical Society, 432*(11), 359-373. doi:10.48550/arXiv.astro-ph/0107201

West, M. J. (1989, Dec. 15). On the morphology of superclusters. *Astrophysical Journal,, vol. 347*, p. 610-626.