Galaxy Morphological Classification: An Analysis of Large Spectral and Photometric Datasets*

Extended Abstract

Son Pham†
CU Boulder
Boulder, CO
son.pham-2@colorado.edu

ABSTRACT

CCS CONCEPTS

• **Mathematics of computing** → *Mathematical analysis*; • **Information systems** → *Data management systems*; • **Computing methodologies** → *Machine learning*; *Distributed computing methodologies*; • **Applied computing** → *Astronomy*.

KEYWORDS

ACM proceedings

ACM Reference Format:
Son Pham. 2022. Galaxy Morphological Classification: An Analysis of Large Spectral and Photometric Datasets: Extended Abstract. In *Proceedings of CSCI 6502 - Big Data Analytics (Spring '22).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.475/123_4

*Produces the permission block, and copyright information
†Note

Spring '22, Mar '22, Boulder , CO, USA

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

1 INTRODUCTION

Big data analytics is transforming the way scientists and astronomers study the universe. With each large scale telescope on the ground or in space, dozens of gigabytes of data is being generated per day - an amount that will take an increasingly amount of time to explore and process. For example, the Hubble Space Telescope generates about 120G of scientific data every week [10]. Sky surveys conducted by telescopes such as the Sloan Digital Sky Survey (SDSS) produces data releases each year that can be as high as 100's of TBs. SDSS provides a wide range of data types, such as optical spectra, infrared spectra, and imaging.

Recent discoveries, such as the presence of over 1000 black holes in the center of our Milky Way, was realized using data from decades ago generated by the Chandra satellite [5]. Scientific and technological advancement in combination led the way to this capability. With the increase in computational performance, astronomers now have the capabilities to explore these large datasets without the need to invest in large ground-based optical telescopes or work in a research lab. As big data analytics continue to grow, these discoveries will become more common as scientists continue to collect and process more of the data that is available. As technological advancements grow, they will have more readily available tools and platforms for their analysis. With the cummilation telescopes and technologies present today, the entire electromagnetic spectrum can be observed within a patch of sky of interest.

SDSS datasets, such provide many different data types, provides the opportunity to explore galaxies, stars, and

quasars too distant to view for amateur astronomers and conduct many different types of analysis on them. The analytics of interest to the authors include the use of optical imaging and spectral data. Legacy imaging generated from prior SDSS programs, if used with machine learning classification techniques, can provide automatic classification of morphological properties of these galaxies. Quatities measured from these images and spectra readings can also provide information such as magnitudes, redshifts, and object classifications. In particular, redshift can be a good indicator of galaxy morphology.

This paper will go over in detail the approach to perform morphological classification of galaxies using these available information from various sources. The main source of interest for the author is SDSS's data, which includes 100's of terabytes of data covering more than one-third of the entire celestial sphere [11].

2 LITERATURE SURVEY

Sky surveys provide a plethora of data for scientists to study and observed. Many different studies were conducted with data collected from satellites and telescopes around the world. The field of morphological classification, in particular, is rich in the research of techniques and methods to analyze these large datasets and help classify galaxies. Galaxy Zoo, a program that spearheaded the galaxy morphological classification manually conducted by volunteer astronomers around the world [1] [4].

G. Martin et al conducted studies on galaxy morphological classification using unsupervised machine learning techniques [2]. The technique the authors proposed does not require classified data, which is beneficial for new surveys that does not come with class labels. In the method proposed, clustering was used to present graphs that represent image patches with similar visual properties, which can be used to group similar galaxies together.

H. Dominguez Sanchez et al. studied machine learning techniques to improve the classification process, particularly in the field of deep learning [3]. This process utilized existing visual classification catalogues with neural network models. Convolutional Neural Networks were trained using two visual classification catalogues: the Galaxy Zoo 2 similar to the current project's were used to train the geometric parameters that makes up a galaxy, and the Nair & Abraham datasets for imaging attributes.

P. H. Barchi et al. presented a paper on comparing various galaxy morphological classification techniques and provided suggestions on how to substantially improve the classification process [6].

Parameter	Description	Unit
Total unique area covered	14,555	sq. deg
Total area of imaging (including overlaps)	31,637	sq. deg
Individual image field size	1361x2048 (0.0337)	pixels (sq. deg)
catalog objects	1,231,051,050	(-)
unique detections	932,891,133	(-)
Median PSF FWHM, r-band	1.3	arcsec
Pixel scale	0.396	arcsec

Table 1: SDSS Dataset Sky Coverage [11]

3 DATASETS

The main dataset source used in this project is the Sloan Digital Sky Survey (SDSS). SDSS provides all the data captured from its instruments and encompasses over one-third of the entire celestial sphere. The latest data release from SDSS is Data Release 17 in December 2021 [11], with an overview shown in Table 1.

The main source used within SDSS was the data collected from their imaging system, which contains objects for about one-third of the entire night sky. Five broad bands were defined and used by SDSS imaging camera, namely "u-g-r-i-z", which are the five rows of Charge Couple Devices (CCD's) used by the imaging camera. The central wavelengths of the filters are shown in Table 2, and a plot of the response curves of the color band is shown in Figure 1, which shows the throughput that defines the photometric system. More information on the camera system can be found on the SDSS imaging website [9].

Although the images is not currently being used in the project, data from the five color bands are. This dataset contained almost half a billion unique objects (stars, galaxies, etc.), with many parameters, though only a selected number of key parameters were used in the machine learning model, shown in Table 3. the celestial right ascension and declination is used to locate

Filter Band	Wavelength (Å)
u	3551
g	4686
r	6166
i	7480
z	8932

Table 2: SDSS Color-band Filters [9]

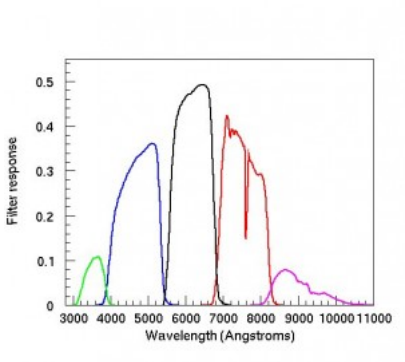

Figure 1: SDSS Imaging Colorbands [9]

the object in the celestial sphere and to cross-match across other data sources. The color bands are used to determine the dominant filter bands present in an object and determine redshift. Adaptive 4th moments of object intensity are measured using a radial Gaussian weight function, teratively adapted to the shape and size of the object. The radius containing the 50% and 90% total Petrosian flux in each band is used to determine the concentration of light for an object, which is the ratio of the radius from the center of an object to the radius containing 50% and 90% of the total Petrosian flux. Reference [12] provides an overview of using Petrosian flux to determine the shape of a galaxy.

Another data source used from the SDSS are the optical spectra data collected by the Extended Baryon Oscillation Spectroscopic Survey (eBOSS) [8]. In essense, the mission of eBOSS is to measure the expansion history of the Universe, focusing on observations of galaxies and quasars, in particularly the rate of change of distances of these objects (redshift). Table 4 lists the parameters used from this dataset. As better understanding of the survey is obtained, additional parameters may be used to increase model fidelity.

The final dataset currently used in the project is the morphological classification dataset from Galaxy Zoo,

Parameter	Description	Unit
Objid	object id	-
ra	Celestial Right Ascension	deg
dec	Celestial Declination	deg
u,g,r,i,z	SDSS color-band filters	Å
mCr4_u/g/r/i/z	Adaptive 4th moments of object intensity	(-)
petroR50_u/g/r/i/z petroR90_u/g/r/i/z	Radius containing 50/90% total Petrosian Flux in each band	arcsec

Table 3: SDSS Photometric Parameters Used

Parameter	Description	Unit
bestobjid	recommended SDSS object match	-
class	best spectroscopic match (star, galaxy, qso)	-
subclass	best spectroscopic subclassification	-
z	best redshift	Å
zerr	error in best redshift	-

Table 4: eBOSS Spectro Parameters Used

which contains classifications for nearly 900,000 galaxies. To better understand details about a galaxy that cannot be measured from current instruments, Galaxy Zoo saught out to study the shape of them which can provide further information about its properties. For example, spiral galaxies typically contains a rotating disk of stars and dust and gas filled with resources for future star formation, whereas elliptical galaxies are more mature and may have finished forming new stars a long time ago. The parameters used from the dataset is listed in Table .

A part of the SDSS instruments suite, the APOGEE-2 Survey contains high-resolution, near-infrared spectrograph that provides spectras across the H-band wavelength regime, which lies between the near-infared 1.51-1.70 micron spectrum, with an approximate resolution of 22,500. Although not currently used in the project, the author anticipate the information it contains from

Parameter	Description	Unit
ra	celestial right ascension	deg
dec	celestial declination	deg
gz_class gz2_class	best class (elliptical, clockwise spiral, anticlockwise spiral, edge-on , star/don't know, or merger)	-

Table 5: Galaxy Zoo Parameters Used

the H-band will benefit the analysis once the initial milestones are completed.

4 APPROACH

4.1 Data Integration

There are many available sources that provide celestial sky survey data. In astronomy, many use a common file transfer type called Flexible Image Transport System (FITS), mainly used for image file transfer, but can also be used to store any type of data that can fit in a multi-dimensional array or table. Typical sky surveys contain equitorial coordinates, defined relative to the celestial sphere, that can be used to identify position of an object in the sky. These surveys also contain object ID's that is used to reference a particular object within a catalog or database. These object ID's are then used as keys to integrate various databases together.

To link the many objects from various data sources, the common method is to obtain the right ascension (RA) and declination (DEC) of the object in the celestial frame. RA is measured from the vernal equinox to the point of interest, going positive east along the celestial equator. The vernal equinox is the intersection of the celestial equator and the ecliptic where the ecliptic crosses the ascending node. The RA varies from 0 to 24 hours, and typically measured in hours-minutes-seconds (HMS). To get RA in decimal degrees, Equation 1 is used.

DEC is the angle from the celestial equator to the point of interest, positive going north (and negative going south). DEC varies from -90 to +90 degrees, with fractions of a degree measured in arcminutes and arcseconds. The notation for DEC is degrees-minutes-seconds, with a full circle providing 360 degrees, each degree having 60 arcminutes and each arcminute having 60 arcseconds. DEC in decimal degrees can be obtained using Equation 2. Some surveys, such as the SuperCOSMOS survey, already has their data in this format, but sources such as the Australia Telescope 20-GHz Survey contains coordinates in the standard form and must be converted.

$$RA_{deg} = 15(hours + \frac{minutes}{60} + \frac{seconds}{3600}) \quad (1)$$

$$DEC_{deg} = degrees + sign(degrees)(\frac{arcminutes}{60} + \frac{arcseconds}{3600}) \quad (2)$$

Once all sources format is converted in to a common decimal degrees format, the datasets can be read and integrated into a common database. To connect objects from different surveys together, we must look at the coordinates of the objects within each catalog and cross-match them to find their closest counterparts, measured by the angular distance between them. Since the position of two given objects are measured as points on a sphere, the great-circle distance must be computed. This can be achieved using the haversine formula, which computes the distance of two points using a set of right ascension and declination angles, as shown in 3, where α and δ are the corresponding RA and DEC of the two points.

$$d_\theta = 2\arcsin\cdot\sqrt{\sin^2\frac{|\delta_1-\delta_2|}{2} + \cos\delta_1\cos\delta_2\sin^2\frac{|\alpha_1-\alpha_2|}{2}} \quad (3)$$

Although this computation is useful in helping integrate multiple sources together, the algorithm does not scale well. Since the algorithm requires the calculation of the distance of each data object from one catalog to each data object in the second catalog, the time complexity is O(n*m), where n and m are the counts of data objects in the catalogs. This is not an issue for catalogs with a few hundred samples, but typical catalogs such as the SDSS contains over a million data points, which will be prohibitively expensive to compute.

An improved algorithm was used for this project's analysis. The first method was to use Python's numpy array structure to automatically loop through a catalog dataset. This provided a speed boost compared to

standard Python standard library, but slowed down significantly when working with datasets greater than 10 million data objects. The second method was to contrain the search to within a given angular radius away from the desired celestial location, which will disregard points located outside of this radius. To find this index value within a catalog, a binary search is performed, where the catalog is repeatedly split in half until the value of interest is found. This provided a boost in performance for small and medium sized datasets, but slows down significantly for datasets with over 100 million data objects. The final method was to use a k-d tree to perform the crossmatching. Similar to the binary search algorithm, the k-d tree divides the k-dimensional space into two parts recursively until each segment is it's own leaf.

Dan Gao et. al provided detailed overview of the k-d tree implementation to astronomy applications [7]. The algorithm requires the kd-tree is built once, then the crossmatching query only need to reference the tree instead of having to scan the entire catalog. Each branch in a tree represents a subvolume of the catalog space, with the intermediate nodes (non-leaf) branching off to exactly two new branches determined by splitting the parent's bounding box. The left new branch contain objects that are less than the splitting value and the right branch contain objects greater than that value.

4.2 Statistical Analysis

One of the major issues in working with large datasets is that the standard analytical evaluation of statistical formulas requires an entire batch of data samples to be stored in memory for computation. Having 100's of thousands of samples can quickly exceed the memory limit of an average laptop or computer. This is especially the case when running analysis on embedded systems where memory is a driving constraint of performance. Stream processing allows a user to run calculations as the data comes in and releases memory that held samples of data from previous use.

Welford's method for statistical analysis allows the use of stream processing to compute some statistical parameters [13]. This method gives accurate estimates of the mean and variance without having to store all the data in memory. The standard process for computing standard deviation is to compute the mean of the data in one pass, then calculate the square deviation of values from the mean in the second pass. In crude methods of numerically computing deviation and means, one can compute the same standard deviation in one pass. Equation 4 shows this method by accumulating the sums of x_i and x_i^2. This subtraction can result in loss of accuracy if the square of the mean is large while the variance is small.

$$\sigma = \sqrt{(n \sum_{1 \le i \le n} x_i^2 - (\sum_{1 \le i \le n} x_i)^2)/n(n-1)} \tag{4}$$

Welford's method simply keeps a running sum of the data, number of samples, and deviation from data collected so far, and the user can view the sample variance and mean at anytime during the computational process to view their progression. Equation 5 show the recurrence formula from Welford which takes into account only the current and previous sample values. Equation 6 shows the computation of the mean, and Equation 7 for the standard deviation.

$$M_{2,n} = M_{2,n-1} + (x_n - \bar{x}_{n-1})(x_n - \bar{x}_n) \tag{5}$$

$$\sigma_n^2 = \frac{M_{2,n}}{n} \tag{6}$$

$$s_n^2 = \frac{M_{2,n}}{n-1} \tag{7}$$

5 RESULTS

5.1 Statistical Analysis

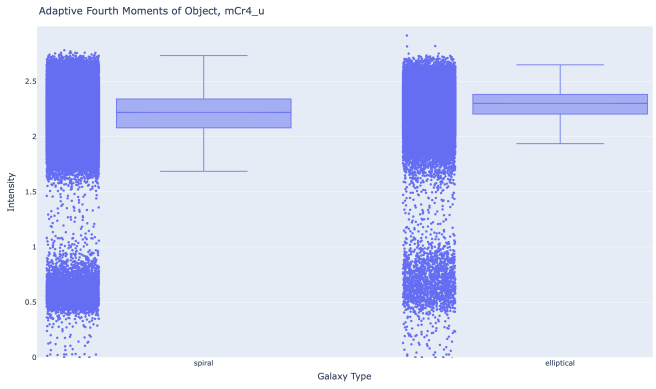

Figure 2: Adaptive 4th Moments - mCr4u

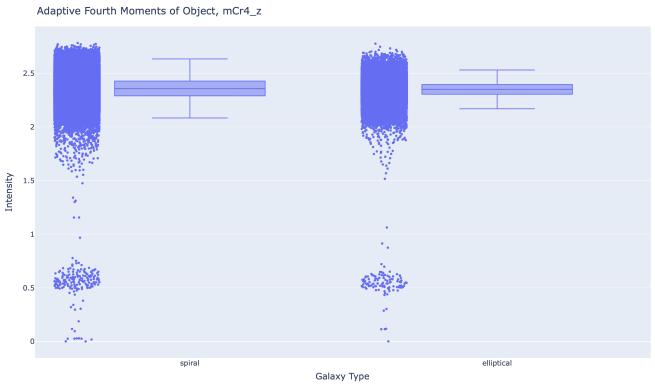

Figure 3: Adaptive 4th Moments - mCr4u

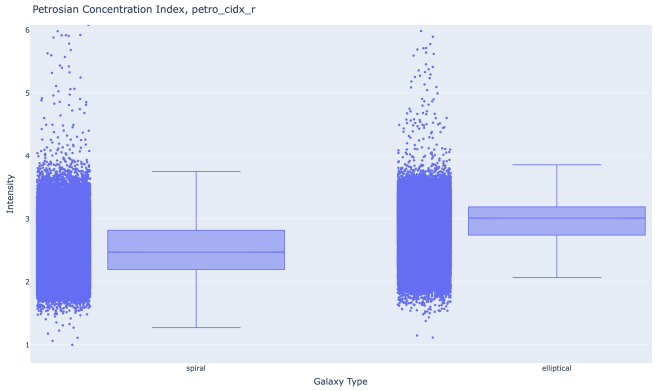

Figure 4: Petrosian Concentration Index, r

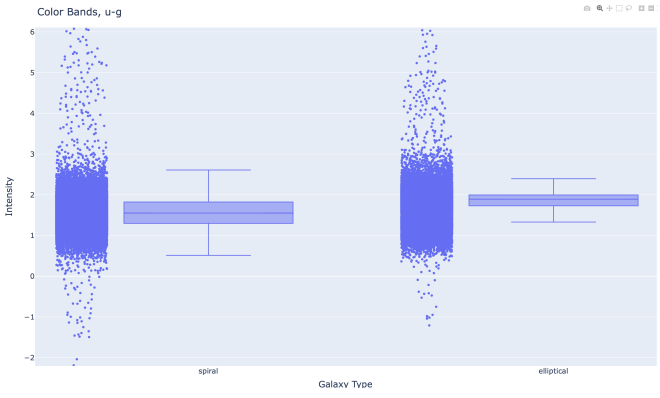

Figure 5: Color Filter - u-g

5.2 Machine Learning

Figure 6: Redshift vs. Right Ascension

# Of Trees	F1 (228K Objects)	Speed (228K)	F1 (100K Objects)	Speed (100K) (S)	F1 (50K Objects)	Speed (50K) (S)	F1 (25K Objects)	Speed (25K) (S)
1	0.7303	9.78	0.7442	4.10	0.7498	1.98	0.7761	0.98
5	0.7805	32.23	0.8044	13.09	0.8073	6.12	0.8252	2.89
10	0.8082	54.23	0.8322	22.30	0.8343	10.08	0.8505	4.85
15	0.8178	71.60	0.8392	29.00	0.8413	13.53	0.8571	6.24
20	0.8203	82.51	0.8385	33.42	0.8417	15.17	0.8595	6.82
30	0.8205	90.19	0.8398	35.61	0.8410	15.89	0.8563	6.98
50	0.8210	89.98	0.8395	35.63	0.8390	15.79	0.8549	6.93

Figure 7: Random Forest Results

# Of Learner	F1 (228K Objects)	Speed (228K) (S)	F1 (100K Objects)	Speed (100K) (S)	F1 (50K Objects)	Speed (50K) (S)	F1 (25K Objects)	Speed (25K) (S)
1	0.7267	2.72	0.7495	1.19	0.7629	0.58	0.7871	0.31
5	0.7431	5.19	0.7688	2.27	0.7896	1.08	0.7944	0.53
10	0.7666	8.46	0.7875	3.50	0.7964	1.71	0.8044	0.82
20	0.7786	14.56	0.7967	6.07	0.8077	2.93	0.8152	1.41
50	0.7912	33.04	0.8093	13.52	0.8171	6.52	0.8370	3.14
100	0.7988	62.84	0.8157	25.77	0.8245	12.74	0.8458	6.01
200	0.8040	121.54	0.8213	49.33	0.8287	24.59	0.8477	11.63
500	0.8078	293.48	0.8257	120.40	0.8366	58.97	0.8507	27.78

Figure 8: AdaBoost Results

# Of Estimator	F1 (228K Objects)	Speed (228K) (S)	F1 (100K Objects)	Speed (100K) (S)	F1 (50K Objects)	Speed (50K) (S)	F1 (25K Objects)	Speed (25K) (S)
1	0.3646	3.81	0.4028	1.62	0.4030	0.80	0.4104	0.38
5	0.7471	10.41	0.7002	4.32	0.7111	2.01	0.7119	0.96
10	0.7804	18.69	0.7824	7.72	0.7692	3.63	0.7956	1.70
20	0.7893	34.94	0.8074	14.27	0.8066	6.54	0.8260	3.08
50	0.8021	83.73	0.8226	33.76	0.8260	16.19	0.8498	7.52
100	0.8107	162.53	0.8293	66.33	0.8360	31.99	0.8599	14.44
200	0.8169	318.45	0.8353	128.69	0.8390	62.66	0.8634	29.11
500	0.8218	774.50	0.8404	316.01	0.8454	151.51	0.8651	70.97

Figure 9: Stochastic Gradient Boost Results

6 DISCUSSIONS

6.1 Results

6.2 Applications

7 CONCLUSION

8 APPENDIX

Honor Code Pledge

# Of Neighbor	F1 (228K Objects)	Speed (228K) (S)	F1 (100K Objects)	Speed (100K) (S)	F1 (50K Objects)	Speed (50K) (S)	F1 (25K Objects)	Speed (25K) (S)
1	0.7060	131.96	0.7303	24.73	0.7249	7.07	0.7295	1.77
5	0.7571	152.52	0.7791	34.61	0.7803	8.81	0.7866	2.07
10	0.7656	152.52	0.7860	34.64	0.7852	9.07	0.7881	2.01
20	0.7735	152.52	0.7919	34.49	0.7930	8.91	0.7973	2.06
30	0.7747	152.55	0.7942	34.54	0.7952	9.05	0.8023	2.06
50	0.7757	152.56	0.7944	34.70	0.7956	8.89	0.8022	2.08

Figure 10: K-Nearest Neighbors Results

# Of Layers	F1 (228K Objects)	Speed (228K) (S)	F1 (100K Objects)	Speed (100K) (S)	F1 (50K Objects)	Speed (50K) (S)	F1 (25K Objects)	Speed (25K) (S)
(1,1)	0.7853	11.27	0.8078	8.70	0.8141	6.42	0.8217	3.20
(2,2)	0.7962	12.48	0.8167	8.14	0.8145	7.07	0.8143	2.04
(5,2)	0.8118	16.81	0.8182	8.96	0.8267	6.62	0.8340	2.52
(5,5)	0.8094	15.97	0.8310	12.60	0.8408	4.57	0.8545	4.12
(2,2,2)	0.7859	33.54	0.8082	15.14	0.8133	10.22	0.8169	3.34
(4,4,4)	0.8099	19.06	0.8318	11.11	0.8380	6.44	0.8386	3.77
(8,8,8)	0.8167	17.56	0.8365	13.78	0.8414	6.55	0.8535	3.49
(2,2,2,2)	0.7976	24.31	0.8142	5.42	0.8178	2.35	0.8298	1.94
(5,5,5,5)	0.8100	32.56	0.8315	11.69	0.8396	6.03	0.8559	3.47
(8,8,8,8)	0.8154	23.34	0.8348	14.87	0.8417	9.08	0.8557	3.29

Figure 11: Multi-Layer Perceptron NN Results

REFERENCES

[1] Chris Lintott et al. 2010. Galaxy Zoo 1: data release of morphological classifications for nearly 900 000 galaxies. *Monthly Notices of the Royal Astronomical Society* 410 (July 2010). doi:10.1111/j.1365-2966.2010.17432.x

[2] G. Martin et al. 2019. Galaxy morphological classification in deep-wide surveys via unsupervised machine learning. *Monthly Notices of the Royal Astronomical Society* 491 (Sept. 2019). doi:10.1093/mnras/stz3006

[3] H. Dominguez Sanchez et al. 2017. Improving galaxy morphologies for SDSS with Deep Learning. *Monthly Notices of the Royal Astronomical Society* 476 (Dec. 2017). doi: 10.1093/mnras/sty338

[4] Kyle W. Willett et al. 2013. Galaxy Zoo 2: detailed morphological classifications for 304 122 galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society* 435 (July 2013). doi:10.1093/mnras/stt1458

[5] NASA/CXC/Columbia Univ./C. Hailey et al. [n.d.]. Sagittarius A* Swarm: Black Hole Bounty Captured in the Milky Way Center. https://chandra.harvard.edu/photo/2018/sgra_swarm/

[6] P. H. Barchi et al. [n.d.]. Machine and Deep Learning Applied to Galaxy Morphology - A Comparative Study. *Astronomy and Computing* ([n. d.]). arXiv:1901.07047v5[astro-ph.IM] 1Nov2019

[7] Dan Gao, Yanxia Zhang, and Yongheng Zhao. [n.d.]. The Application of kd-tree in Astronomy. *Astronomical Data Analysis Software and Systems XVII* 394 ([n. d.]).

[8] SDSS. [n.d.]. The Extended Baryon Oscillation Spectroscopic Survey (eBOSS). https://www.sdss.org/surveys/eboss/

[9] SDSS. 2021. SDSS Imaging Camera. https://www.sdss.org/instruments/camera/

[10] Nola Taylor Tillman. 2022. Hubble Space Telescope: Pictures, facts and history. (2022).

[11] Abdurro uf et al. 2022. The Seventeenth data release of the Sloan Digital Sky Surveys: Complete Release of MaNGA, MaStar and APOGEE-2 DATA. *The Astrophysical Journal* 259, 2 (March 2022). doi:10.3847/1538-4365/ac4414

[12] St. Lawrence University. 2021. Galaxy Photometry. http://myslu.stlawu.edu/~aodo/astronomy/ALFALFA/Astronomy%20Basics/GalaxyPhotometry.pdf

[13] B. P. Welford. 1962. Note on a Method for Calculating Corrected Sums of Squares and Products. *J. ACM* 4, 3 (Aug. 1962). https://www.jstor.org/stable/1266577