Morphological Classification of Galaxies using Machine Learning Algorithms

by

Madala Vikas
Roll No. IMS21106

SCHOOL OF DATA SCIENCE
INDIAN INSTITUTE OF SCIENCE EDUCATION AND RESEARCH
THIRUVANANTHAPURAM - 695 551, INDIA

May 2024

DECLARATION

I, **Madala Vikas (Roll No: IMS21106)**, hereby declare that, this report entitled **"Morphological Classification of Galaxies using Machine Learning Algorithms"** submitted to Indian Institute of Science Education and Research Thiruvananthapuram towards the partial requirement in **DSC 325**, is an original work carried out by me under the supervision of **Dr Pranshu Mandal**. I have sincerely tried to uphold academic ethics and honesty. Whenever a piece of external information or statement or result is used then, that has been duly acknowledged and cited.

Thiruvananthapuram - 695 551 **Madala Vikas**

May 2024

ACKNOWLEDGEMENT

I thank everyone who has assisted me in seeing this project through to its completion. I would like to first express my profound gratitude and deepest regards to Dr Pranshu Mandal, IISER Thiruvananthapuram, and sincerely wish to acknowledge his vision, guidance, valuable feedback and constant support throughout the duration of this project.

I am indebted to Dr Pranshu Mandal for their steadfast encouragement and time. I am lastly grateful to the Indian Institute of Science Education and Research Thiruvananthapuram for providing the necessary resources and facilities to complete this project to the best of my ability.

Thiruvananthapuram - 695 551 **Madala Vikas**

May 2024

ABSTRACT

Name of the student: **Madala Vikas** Roll No: **IMS21106**

Course for which submitted: **DSC 325** Department: **School of Data Sience**

Project title: **Morphological Classification of Galaxies using Machine Learning Algorithms**

Project supervisor: **Dr Pranshu Mandal**

Date of Report submission: **May 2024**

The project aims to classify the Galaxies based on their morphological properties by making use of Machine Learning techniques. This project presents a new method for classifying galaxies using the EFIGI attributes data. For this purpose, we have used various feature selection and geneartion methods alongwith a recent algorithm called SMOTE, which is an oversampling method. Decision Tree classifier, Random Forest classifier, Gradient Boosting classifier have been considered for the classification. We have also considered the classes in two different scenarios, one with five classes and the other with 18 classes which are the subclasses of the previously mentioned five classes. With these methods, we have acquired a mean cross-validation score of 87% on the five classes scenario and 42% on the 18 classes.

Keywords:

Morphology of galaxies, ANOVA(Analysis of Variance), Cross-Validation(CV)
SMOTE(Synthetic Minority Over-sampling Technique)
EFIGI(Extraction de Formes Idealisées de Galaxies en Imagerie)

Introduction

In the observable universe there are about 200 billion to 2 trillion galaxies. To find out the reasons behind the nature, organization and history of the universe, it is very important to study about the galaxies. Especially studying the morphology of the galaxies, helps us to understand evolutionary process, which means, studying the morphology of the galaxies provides the knowledge about the future happenings. For this, one of the first widely used system was Hubble's "tuning fork" (1936), which classifies the galaxies into elliptical, lenticular, spiral or irregular galaxies. In 1959, Revised Hubble System(RHS), has addressed the intermediate stages from "-6 to 10".

To classify the galaxies based on their morphological properties, we have used the modern Machine Learning tools, with which we aim toprovide a tool for the researchers. For this purpose, we have selected the EFIGI Attributes dataset from the Astronomica website and is publicly available. We have used various feature selection methods like ANOVA, SMOTE and have used Random Forest, Decision Tree, and ensemble model Gradient Bossting Classifiers for classification. For the model evaluation, we have considered Accuracy score along with other evaluation metrics like F1 Score to address the imbalance in the data.

About the Dataset

The EFIGI Attributes dataset is comprised of 4458 nearest galaxies with the morphological details. It has 51 feature columns along with the galaxies unique ID column. The EFIGI attributes can be broadly divided into six categories:

1. Appearance: inclination/elongation
2. Environment: multiplicity, contamination
3. Bulge: Bulge/Total ratio
4. Spiral arm properties: arm strength, arm curvature, rotation
5. Textural aspect: visible dust, dust dispersion, flocculence, hot spots
6. Dynamical features: bar length, inner ring, outer ring, pseudo-ring, perturbation

A scale with five steps(0, 0.25, 0.5, 0.75, 1), has been chosen to indicate the strength of an attribute for a given galaxy. Upper and lower limits have been estimated for each attribute of each galaxy with a 70% confidence interval. the attribute "T" is the target variable. We have considered the main 16 main attributes dropping the rest of the attributes which are the upper and lower limit estimates. From these 16 attributes, we then selected the attributes that contribute the most towards the class of the galaxy using various feature selection methods and generation methods. The data was highly imbalanced, and we have taken respective measures for addressing this issue as discussed in the preprocessing section.

Exploratory Data Analysis

There are a total of 18 galaxies in the dataset which are denoted from "-6 to 11". We calculated the total number galaxies for each class and have plotted the results using histograms as shown in Figure.1. The imbalance can be clearly seen from the plot and it is mainly because of the classes that belong to Spiral class.

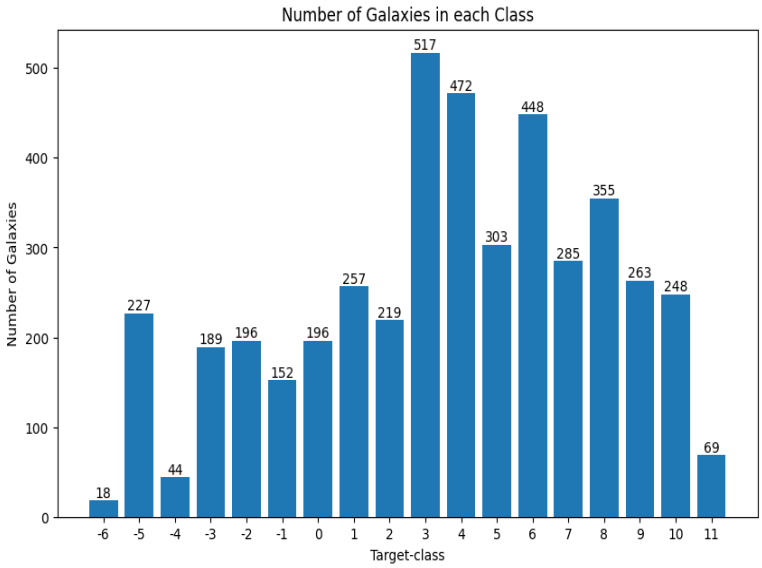

Figure 1: Class distribution of the galaxies

Figure.2 displays the boxplots of the attributes where in each row, an attribute along with it it's minimum and maximum limit estimates are plotted to better understand the distributions of all the attributes which plays a key role in assessing the output.

Also, to study the correlation between the attributes we plotted the heatmap as Figure.3. The dark blue tridiagonal elements in the plot represent the high correla-

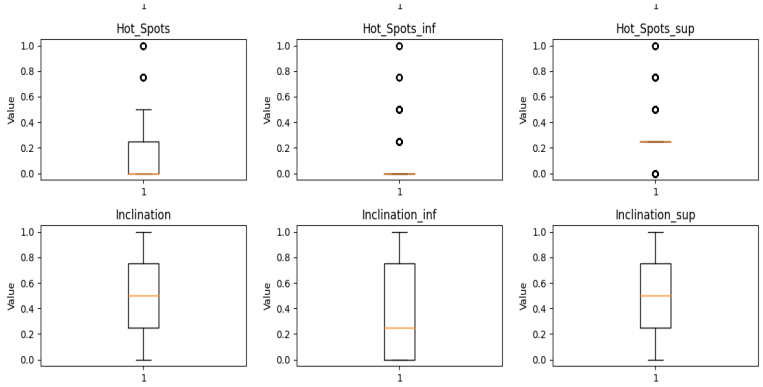

Figure 2: Feature distribution in boxplots

tion between a feature and it's upper and lowe limit estimates.

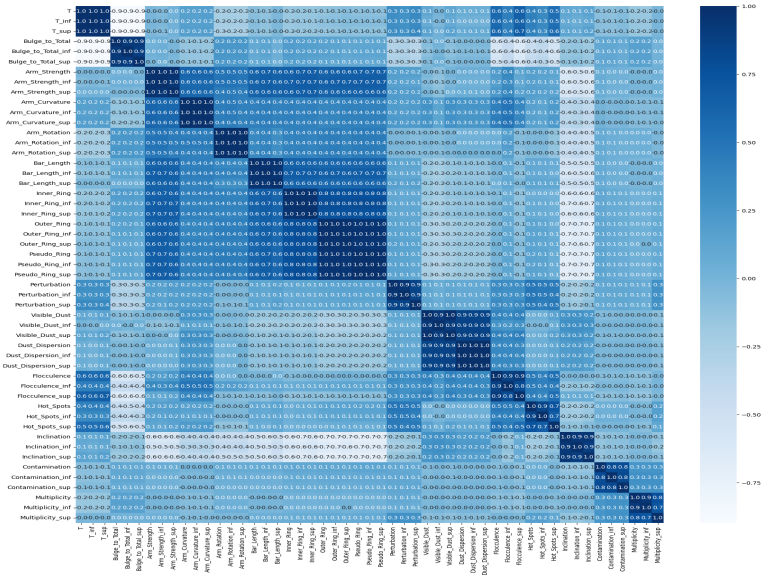

Figure 3: Correlation map

Figure 4 depicts the overview of our project.

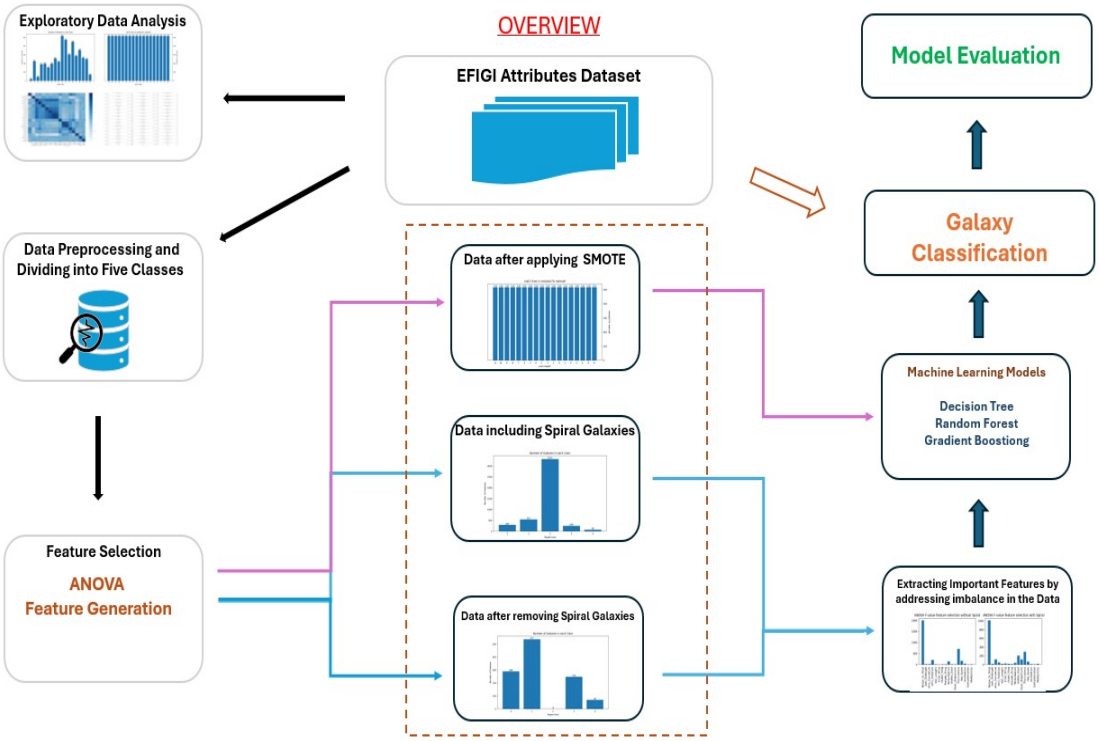

Figure 4: Overview

Preprocessing

The galaxies can be broadly divided into five classes: *Elliptical*, *Lenticular*, *Spiral*, *Irregular* and *Dwarf*. We converted the 18 classes into these five classes for easy interpretation. To get the most contributing features, we used ANOVA F-value method. The data was highly imbalanced as 70% of the data was comprised of *Spiral* galaxies. Noting this, we have applied ANOVA method once with all the data including spiral

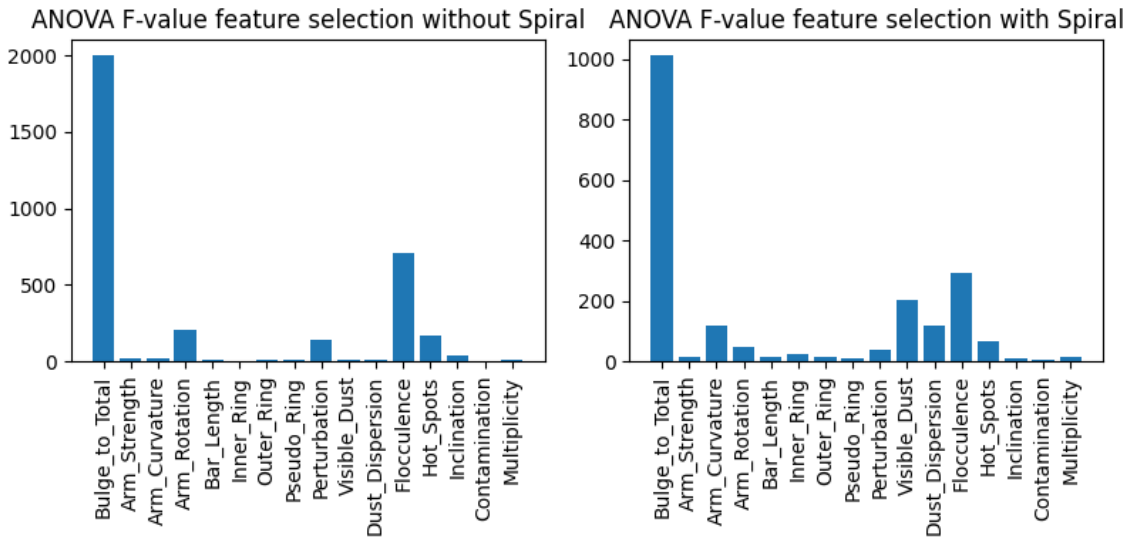

Figure 5: ANOVA with and without spiral class data

galaxies data. By doing this, we tried to find out the main features that contribute of the rest of the galaxy classes and to avoid the bias toward the spiral galaxies. Figure.5 shows the ANOVA F-value plot with and without including the spiral class data. From the plots, we have considered *Bulge to Total, Arm Curvature, Arm Rotation,Perturbation, Flocculence, Hotspots* as the most contributing attributes.

The attribute **Hotspots** is excluded from the attributes that are selected from the ANOVA F-test. The attribute has not significantly contributed to explaining the variance in the target variable or distinguishing between different classes in the dataset. The low variance of the feature could be considered as one of the reason for excluding the attribute. It possessed multicollinearity as well.

We created a new multiplicative feature by feature engineering using the arm curvature and the arm rotation features, as, the quality of the features contributes to the better performance of the model, as, such features help to reduce overfitting.

To address the imbalance in the data, We have also used SMOTE algorithm. This helps to balance the data by adding more instances of the classes with less data and results in the new version of data where each class will have exactly the same number of instances. The below plot is the result of the SMOTE algorithm's application on our dataset.

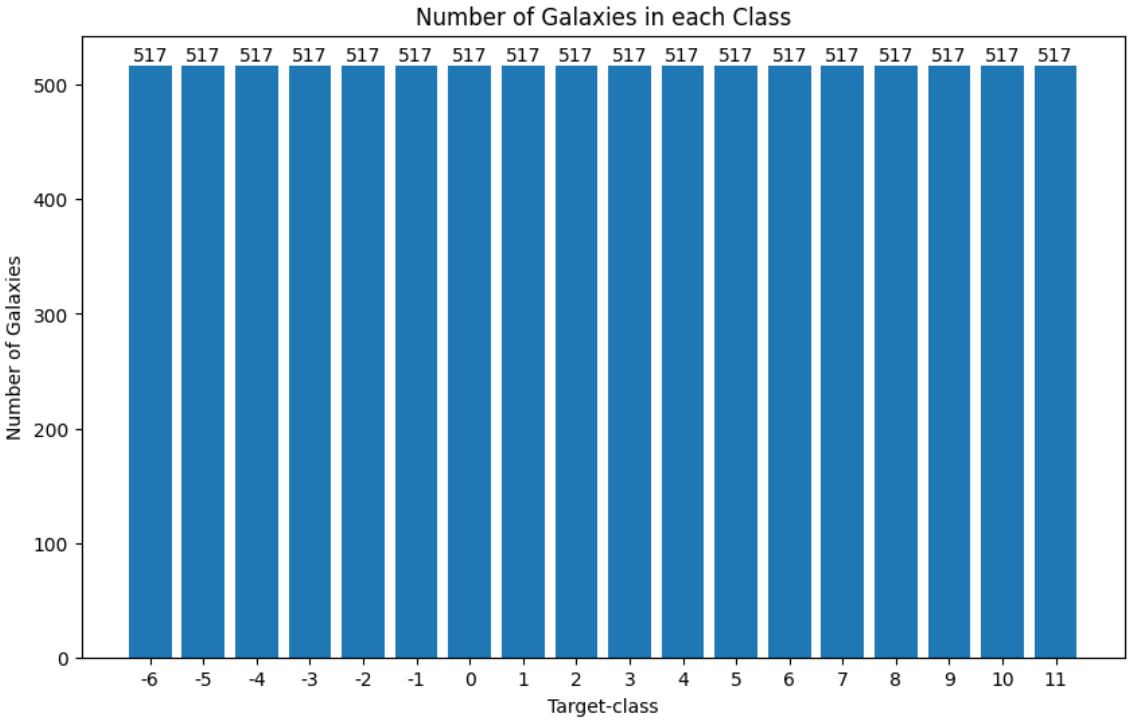

Figure 6: Result of SMOTE algorithm

Methodology

Finally, we have selected the five feature columns and divided the dataset into train data and test data. We use Decision Tree Claasifier, Random Forest Clas-

sifier and Gradient Boosting Classifier for classification. Then we performed 5-fold cross-validation on each classifier and depicted the values in Table 1 and Table 2.

The main reason to choose Cross-validation is that, it assesses a model's performance by splitting data into subsets for training and validation. It aids in evaluating model robustness, reducing overfitting, and providing a reliable estimate of its accuracy. The *cross-validation score* averages performance metrics, is a good evaluation measure of model effectiveness.*Mean CV accuracy* is the average accuracy of a model across all cross-validation folds. It provides a consolidated measure to generalize on the new data. The following are the matematical equations for the Accuracy, Precision, Recall evaluation metrics.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

F1 score is a metric that combines precision and recall into a single value, providing a balanced measure of a model's performance. It considers both false positives and false negatives. A higher F1 score indicates better overall model performance. When data is highly imbalanced F1 score is considered as an essential evaluation metric.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Result and Analysis

Table 1: Cross-validation scores without SMOTE

Classifier	**Mean CV score (5 classes)**	**Mean CV score (18 classes)**
Decision Tree	0.8633	0.3506
Random Forest	0.8681	0.3551
Gradient Boosting	0.8716	0.3802

Table 2: Cross-validation scores with SMOTE

Classifier	**Mean CV score (5 classes)**	**Mean CV score (18 classes)**
Decision Tree	0.8714	0.4218
Random Forest	0.8723	0.4280
Gradient Boosting	0.8686	0.4193

Table 3: Classification Report for Decision Tree Classifier (5 classes)

Class	**Precision**	**Recall**	**F1-Score**	**Support**
0	0.76	0.95	0.85	657
1	0.86	0.68	0.76	683
2	0.93	0.82	0.87	665
3	0.89	0.94	0.92	660
4	0.94	0.97	0.96	650
Accuracy			0.87	3315
Macro Avg	0.88	0.87	0.87	3315
Weighted Avg	0.88	0.87	0.87	3315

Table 1 depicts the mean cross-validation scores without using SMOTE algorithm for five classes and eighteen classes, with highest being the Gradient boosting classifier for both the cases. In case of SMOTE application, the scores have increased for Decision Tree and Random forest classifiers and slightly decreased for Gradient boosting classifier in case of five classes, but, for eighteen classes, the scores have

Table 4: Classification Report for Decision Tree Classifier (18 classes)

Class	**Precision**	**Recall**	**F1-Score**	**Support**
-6	0.65	0.46	0.53	101
-5	0.32	0.82	0.47	108
-4	0.64	0.25	0.35	114
-3	0.29	0.02	0.04	101
-2	0.37	0.65	0.47	110
-1	0.54	0.30	0.38	101
0	0.37	0.39	0.38	95
1	0.37	0.34	0.35	110
2	0.42	0.54	0.48	112
3	0.32	0.36	0.34	98
4	0.25	0.12	0.16	107
5	0.32	0.32	0.32	90
6	0.44	0.63	0.52	115
7	0.43	0.15	0.22	110
8	0.34	0.35	0.34	104
9	0.38	0.26	0.31	101
10	0.53	0.83	0.65	94
11	0.89	0.93	0.91	91
Accuracy			0.43	1862
Macro Avg	0.44	0.43	0.40	1862
Weighted Avg	0.43	0.43	0.40	1862

increased for all the classifiers with Random forest being the highest and the rest of the classifiers have also performed equally well. This is a good score for the 18 class scenario.

Table 3 represents the classification report for Decision Tree classifier on five classes after using SMOTE. The F1 score for the class 1 is low compared to the other classes and could be the reason for lower overall performance. Table 4 represents the classification report for Decision Tree classifier on 18 classes after using SMOTE. Similar to the five class scenario, some of the classes have very less F1 scores resulting

in a lower overall accuracy values.

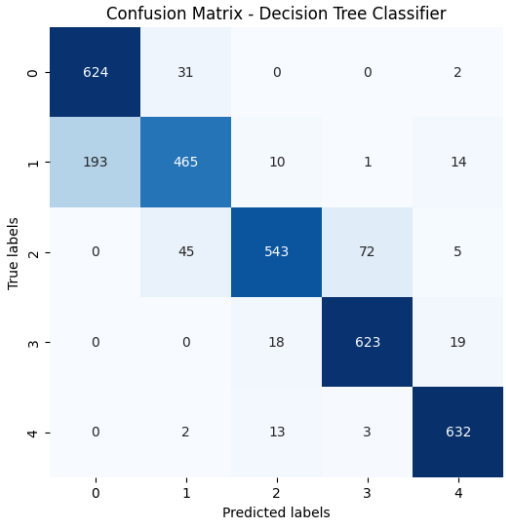

Figure 7: Confusion matrix using SMOTE (5 classes)

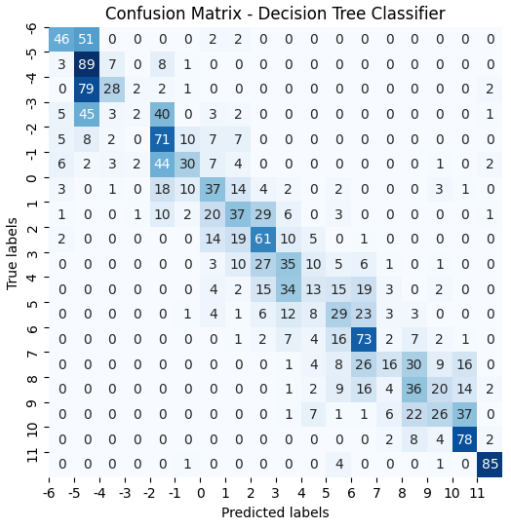

Figure 8: Confusion matrix using SMOTE (18 classes)

Figure 7 and Figure 8 are the confusion matrices corresponding to the above classification report part. It can been seen that the model was not able to classify some of the classes resulting in a poor performance.

The scores can be increased by using even advanced models and gathering more data will also prove beneficial to address the imbalance in the dataset.

Future Research

To improve the efficiency, our future interests are exploring Deep Learning models and working with advanced models on visual data using computer vision and image processing techniques.

Data and Code Availability

Data is publicly available on the EFIGI website and Code has been uploaded in the github.

Bibliography

Baillard, A. et al. "The EFIGI catalogue of 4458 nearby galaxies with detailed morphology". In: *Astronomy amp; Astrophysics* 532 (July 2011), A74. ISSN: 1432-0746. DOI: 10.1051/0004-6361/201016423. URL: http://dx.doi.org/10.1051/0004-6361/201016423.

Fraix-Burnet, D. "Machine learning and galaxy morphology: for what purpose?" In: *Monthly Notices of the Royal Astronomical Society* 523.3 (June 2023), pp. 3974–3990. ISSN: 1365-2966. DOI: 10.1093/mnras/stad1654. URL: http://dx.doi.org/10.1093/mnras/stad1654.

Guruprasad, Anusha. *Galaxy Classification: A machine learning approach for classifying shapes using numerical data*. 2023. arXiv: 2312.00184 [cs.CV].