Astronomical databases and astrostatistics in the era of big data

⋆ **Maria Luísa Gomes Buzzo** ⋆

Term Project: Low surface brightness galaxies within the S-PLUS catalogs

1 Abstract

The detection of low surface brightness galaxies (LSBGs) in imaging surveys is problematic due to the lack of depth, and current instrumentation and pipelines that are not focused on these types of objects. This project aims to use innovative machine learning techniques to automatically detect LSBGs in the S-PLUS survey, thus increasing their number and providing interesting candidates for spectroscopic follow-up. Such techniques can be applied both in catalogs and images, and expanded to future surveys such as that of the Vera Rubin Telescope.

2 Introduction

Since the time when Edwin Hubble gave his great contributions to Modern Astronomy, our knowledge of the field of galaxy formation and evolution has grown significantly. From the Milky Way itself to the farthest galaxy ever observed, we know the most likely scenarios to form galaxies and how these evolve in different environments. We know their chemical enrichment processes, the stellar populations of the different galaxy types and how these evolve with time from the most irregular morphologies at high-redshift to the known morphological types observed today, such as ellipticals and spirals. However, we know as much as the current instrumentation allows us to. For example, the majority of the existing telescopes/surveys have a limiting magnitude of 23 mag/arcsec^2, meaning that most objects fainter than this threshold have remained unknown for quite a long time. This is due to the lack of depth of the surveys, partially, but also to the lack of a pipeline focused on the identification and study of the low-surface brightness universe. Simulations show that low-surface brightness galaxies with $M_\star > 10^7 M_\odot$ contribute to 85% of the local number density of galaxies, but only to 11% of the local mass density and 10% of the local luminosity density [Martin et al., 2019]. However, although being the most common objects in the universe, our current knowledge of galaxies' evolution is mostly based on what is known about the high-surface brightness universe, which could lead to serious biases in our interpretation and understanding of our universe. Using large-area imaging surveys, such as the Dark Energy Survey and the Legacy Survey [Zaritsky et al., 2019, Beasley et al., 2015, Tanoglidis et al., 2021] have found these objects all around the entire sky, which was then followed by important follow-up studies, such as Barbosa et al. [2020], that was able to find and analyse 100 UDGs from the SMUDGES project [Zaritsky et al., 2019] in the S-PLUS survey, unravelling the stellar populations and main properties of these objects. Similarly to these other works, we plan on systematically finding and measuring low surface brightness galaxies using the S-PLUS catalogs. To do so, we will use a sample of known LSBGs that were measured with S-PLUS to train a model using their main properties, looking for them in other regions of the sky, not yet covered by other surveys.

3 Methodology

We gathered a sample of LSBGs found by Zaritsky et al. [2019] in the STRIPE82 region and by Venhola et al. [2017] in the Fornax cluster, as well as a sample of LSBGs observed with the Hyper Suprime Cam in the Subaru telescope [Greco et al., 2018]. Out of the these catalogs, only 122 LSBGs were present in the S-PLUS catalogs, and thus compose the sample that we will use to train and test a model to detect new LSBGs based on their main properties, i.e. surface brightness, magnitude, kron radius, FWHM, ISOarea and others. The distribution of these parameters for our sample -split in STRIPE82, Fornax and SuprimeCam- is shown in Fig. 1.

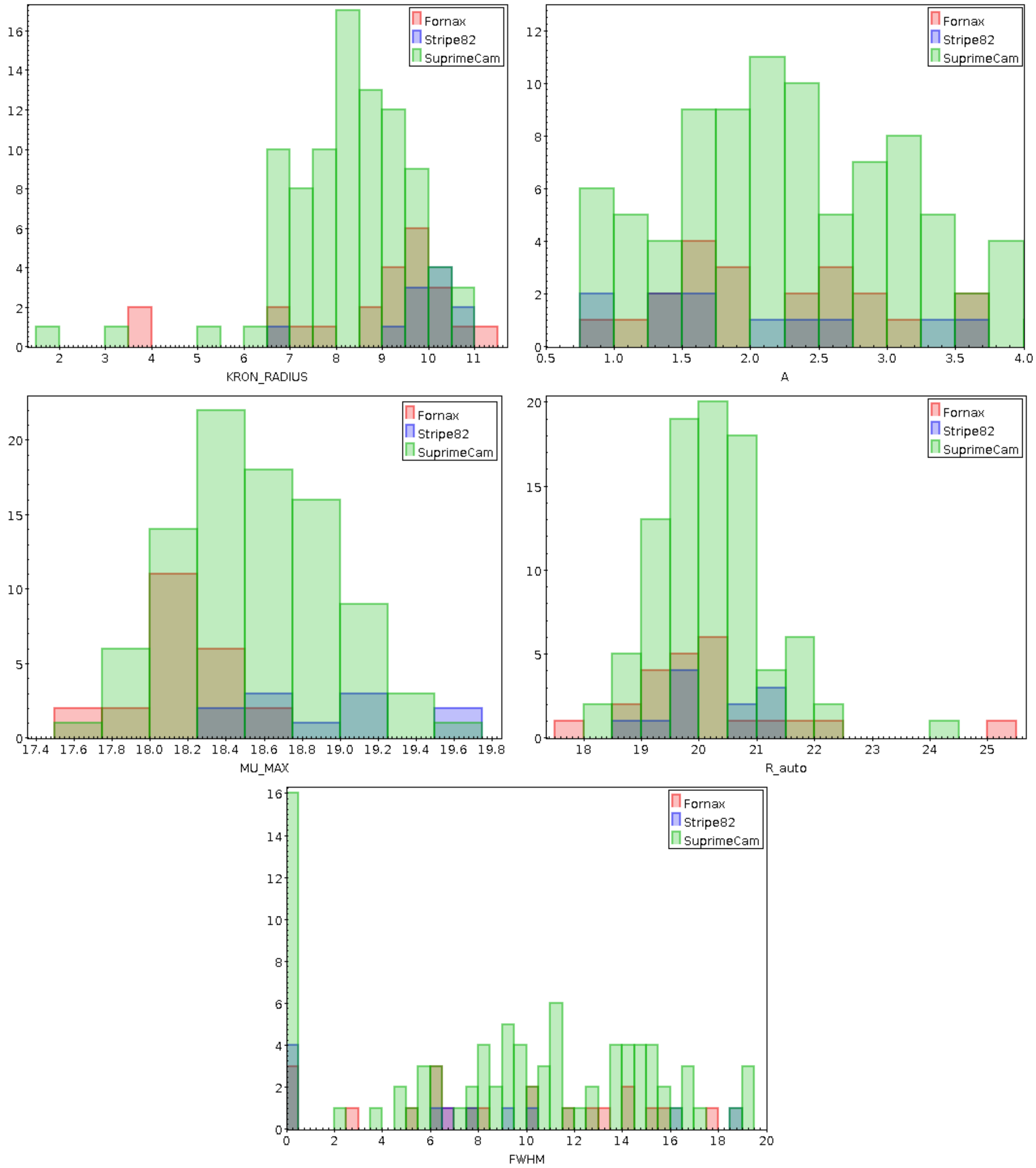

Figure 1: Distribution of key physical parameters of our sample of known LSBGs observed with S-PLUS.

By looking at the range of these features, we downloaded the S-PLUS data that was in the same range so we could apply the classification. This downloaded sample was used to understand which type of object the could be confused with LSBGs. The downloaded data was on the range:

$$6.0 < KRON\_RADIUS < 12.0$$

$$0.5 < A < 4.0$$

$$17.0 < MU\_MAX < 18.5$$

$$18.0 < R\_auto < 23.0$$

$$0.0 < FWHM < 20$$

Additionally, taking our sample of bonafide LSBGs as reference, we considered that the data should be detected in more than 7 bands (nDet_auto $>$7), and not be in the area of any bright star (BrightstarFlag $=$ 0). The entire query to download these data is shown below.

```
SELECT det.ID, det.ra, det.dec, det.CLASS_STAR, det.ELLIPTICITY, det.ELONGATION,
↪  det.FLUX_RADIUS, det.FWHM, det.ISOarea, det.KRON_RADIUS, det.MU_MAX, det.A, u.u_auto,
↪  u.e_u_auto, J0378.J0378_auto, J0378.e_J0378_auto, J0395.J0395_auto,
↪  J0395.e_J0395_auto, J0410.J0410_auto, J0410.e_J0410_auto, J0430.J0430_auto,
↪  J0430.e_J0430_auto, g.g_auto, g.e_g_auto, J0515.J0515_auto, J0515.e_J0515_auto,
↪  r.r_auto, r.e_r_auto, r.r_aper_6, r.r_petro, J0660.J0660_auto, J0660.e_J0660_auto,
↪  i.i_auto, i.e_i_auto,J0861.J0861_auto,J0861.e_J0861_auto, z.z_auto, z.e_z_auto

FROM idr3.detection_image AS det
JOIN idr3.u_band AS u ON (u.ID = det.ID)
JOIN idr3.j0378_band AS J0378 ON (J0378.ID = det.ID)
JOIN idr3.j0395_band AS J0395 ON (J0395.ID = det.ID)
JOIN idr3.j0410_band AS J0410 ON (J0410.ID = det.ID)
JOIN idr3.j0430_band AS J0430 ON (J0430.ID = det.ID)
JOIN idr3.g_band AS g ON (g.ID = det.ID)
JOIN idr3.j0515_band AS J0515 ON (J0515.ID = det.ID)
JOIN idr3.r_band AS r ON (r.ID = det.ID)
JOIN idr3.j0660_band AS J0660 ON (J0660.ID = det.ID)
JOIN idr3.i_band AS i ON (i.ID = det.ID)
JOIN idr3.j0861_band AS J0861 ON (J0861.ID = det.ID)
JOIN idr3.z_band AS z ON (z.ID = det.ID)
JOIN idr3_vacs.masks AS mask ON (mask.ID = det.ID)
JOIN idr3_vacs.star_galaxy_quasar AS sgq ON (sgq.ID = det.ID)
JOIN idr3_vacs.masks AS msk ON (msk.ID = det.ID)

WHERE det.field = 'SPLUS-s27s34' AND sgq.Class = 2 AND mask.BrightstarFlag = 0 AND
↪  det.KRON_RADIUS > 6.5 AND det.KRON_RADIUS < 11.5 AND det.A > 0.5 AND det.A < 4 AND
↪  det.mu_max > 17 AND det.mu_max<18.5 AND r.r_auto > 17.5 AND r.r_auto < 22.5 AND
↪  det.nDet_auto > 7
```

The distribution of the downloaded data is shown in Fig. 2. For this first test, this sample was exclusively for tile SPLUS-s27s34, the central tile of the Fornax cluster.

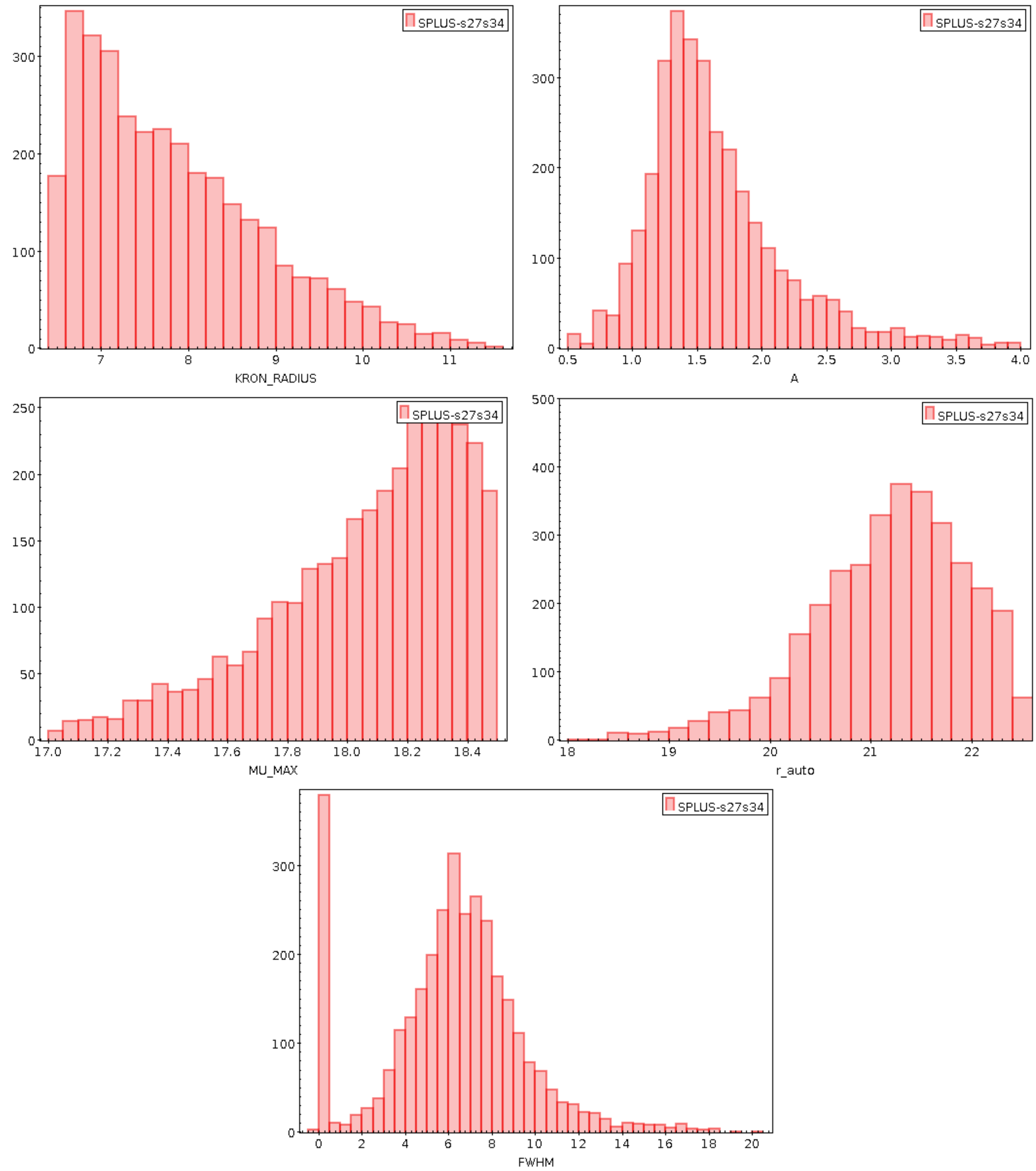

Figure 2: Distribution of physical parameters from the sample of S-PLUS objects recovered using the previous query. These are the full set of objects, i.e. containing those that are yet to be classified as LSBGs/non-LSBGs.

By inspecting the data, we see some misdetections and, moreover, that high-redshift galaxies/red-compact galaxies also occupy this range of the parameter space, so the task of separating LSBGs is not trivial.

4 Results

To start our tests, we gathered our training and testing samples. This was done by joining our sample of known LSBGs (122 galaxies) with a sample of 120 confirmed non-LSBGs. In Fig. 3, we show the plots of the features that will be used to perform the classification, where red points represent the known LSBGs, blue points represent the compact/high-redshift galaxies and green points show the misdetections. From the plots in Fig. 3, we can already see that the LSBGs stand out from the other classes, which could be a good indication that a binary classification is up to the task.

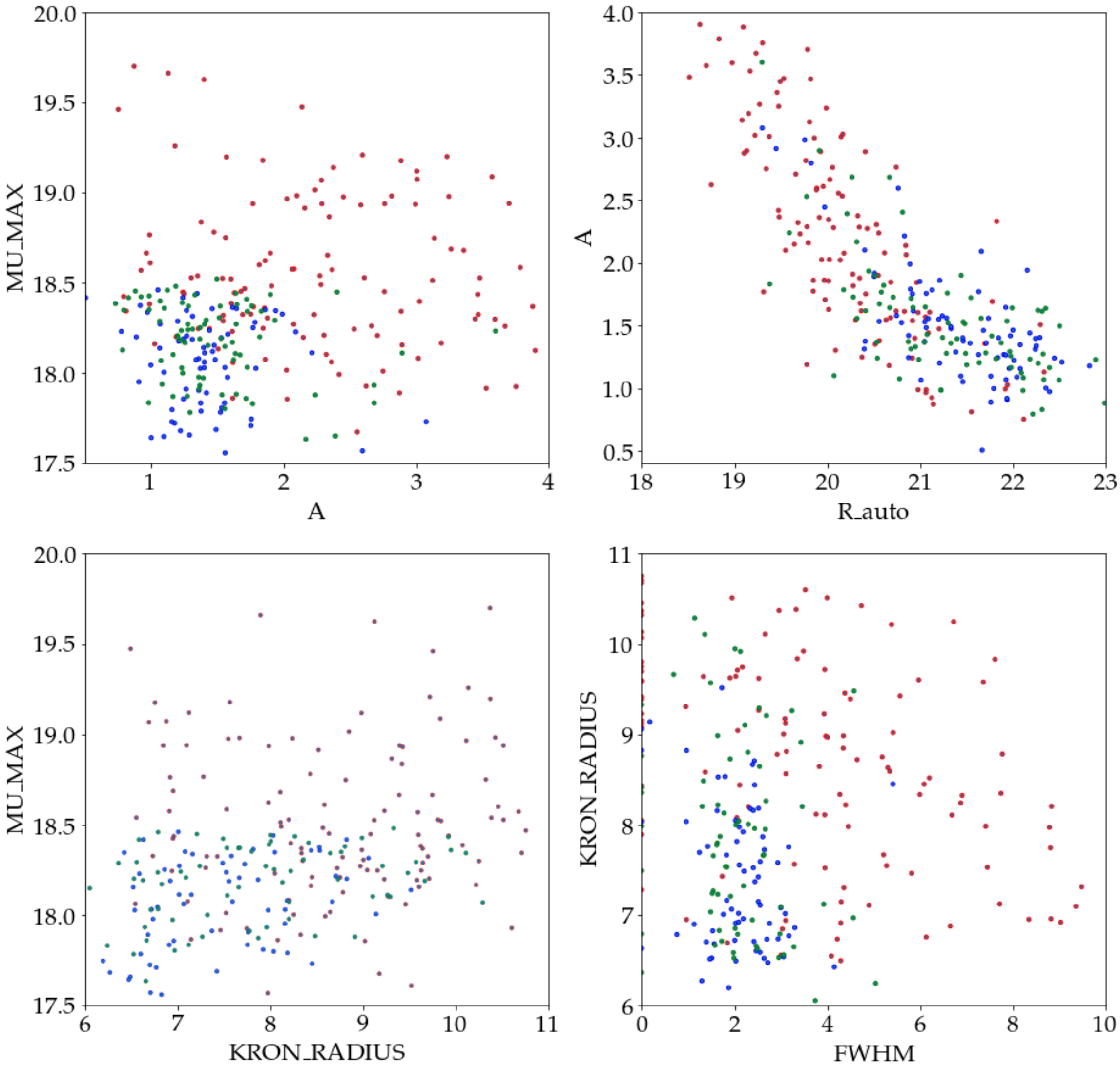

Figure 3: Plots of key physical parameters of our training and testing samples, including bonafide LSBGs (red), red/compact/high-redshift galaxies (blue) and misdetections (green).

We applied the Gaussian Mixture Model technique to each of the physical parameters, allowing from 1 to 3 components. The idea is to identify which features are best to provide a separation between classes. The results of these tests are shown in Fig. 4.

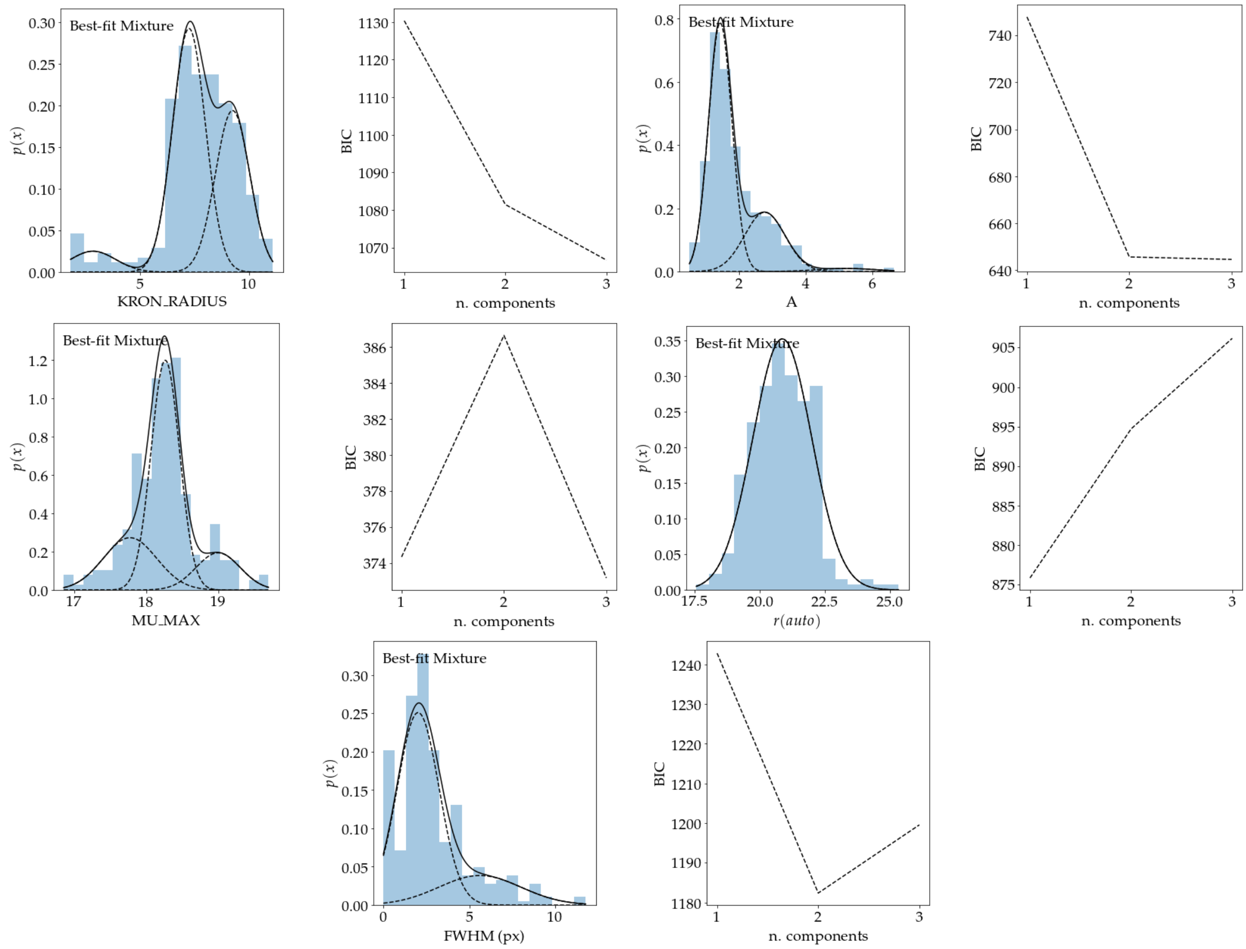

Figure 4: Gaussian Mixture Model results ran on each feature of our training and testing sample. BIC results show how many components are best to explain the data.

From Fig. 4, we can see that KRON_RADIUS, A, MU_MAX and FWHM are good features to train our classification method, because the machine is able to identify more than one component in their distributions. For the r-band magnitude, that is not the case though, and the data is better fitted by a single gaussian. Particularly, we believe that the MU_MAX and KRON_RADIUS features are very promising classifiers to LSBGs, given that these galaxies are fainter and more extended than the other classes, and these features reflect that. Based on all of the previous results, we decided to use four physical parameters to train our classification methods: **KRON_RADIUS, MU_MAX, FWHM and A**.

We used a supervised learning classification with different classification methods, such as Gaussian Naive Bayes (GaussianNB), Quadratic Discriminant Analysis (QDA), K Neighbors Classifier (KNN), Decision Tree Classifier (DT), Gaussian Mixture Models (GMM), and Logistic Regression (LR). We split the data into training (70%) and testing (30%) samples.

In fig. 5, we show the classification of test sample (30% of the total sample) colored by their known classification in the left. On the right, we compare the diagnostic ability of each classification method based on the ROC curves. We can see that all classification methods perform similarly, except for the Decision Tree (purple curve) classifier, that clearly an inferior performance.

Among all the good classifiers, the KNN has a slightly better performance. We note that there is a lot of room for improvement, because the rate of false positives is still high, but we believe that this is a good start and when gathering more data to train our models, these results might significantly improve.

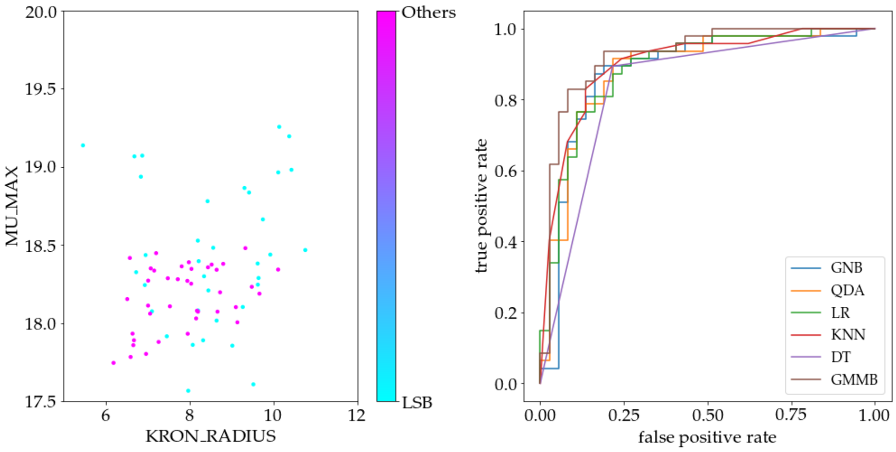

Figure 5: Left: distribution of KRON_RADIUS and MU_MAX as colored by their known classification as LSBGs or others. Right: ROC curves comparing the different classification methods.

After training and testing our method, we applied this classification to an entire tile of SPLUS (SPLUS-s27s34), in the objects that were in the same range as the LSBGs. This sample was composed of 3275 objects. After applying our methods, we verify the probability that each object is labelled as an LSBG, for each method. THe result is shown in Fig. 6. Again, we can see that the classification with the decision tree method is not working well. However, all the others show a sensible distribution, revealing that the objects with highest probabilities of being LSBGs are the ones with highest KRON_RADIUS, while objects with smaller MU_MAX and KRON_RADIUS have very small probabilities of being LSBGs.

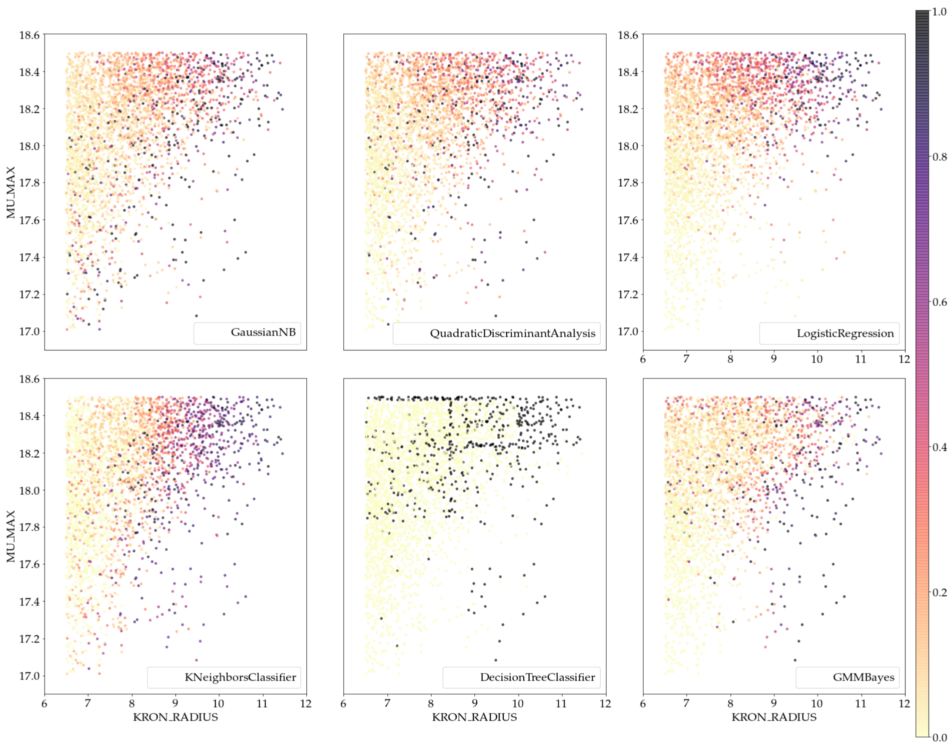

Figure 6: Objects recovered probabilities of being LSBGs using different classification methods.

Selecting all objects that have prob(LSBs) > 0.8, we end up with a sample of 124 galaxies only. We visually inspected these objects and we report 35 new LSBGs in this tile. Some examples are shown in Fig. 7.

Figure 7: Examples of new LSBGs found with S-PLUS

.

5 Conclusion

In this work, we have applied machine learning classification techniques to systematically find low surface brightness galaxies within the S-PLUS survey. Using key features of LSBGs such as surface brightness and radius, we were able to separate objects in two classes: LSBGs or non-LSBGs. The initial sample contained 122 LSBGs and 120 non-LSBGs, further split into training (70%) and test (30%) samples. We report that the best classification was provided by the KNN method, even though the GNB,GMMB,LR and QDA methods also provided good results. We then applied the technique to find LSBGs in the central tile of the Fornax cluster, SPLUS-s27s34. We found 35 new LSBG amongst the 124 candidates with probability higher than 0.8.

Our method was impacted by the quite modest total number of known LSBGs to train and test the classifiers. Nevertheless, we still find very promising results, leading to hundreds of new LSBG candidates in SPLUS, helping to overcome the difficulties implied by a shallow depth.

Increasing the number of LSBGs in the training sample has the potential of providing even better classification and potentially lead to catalogs of thousands of LSBGs within the survey. As additional future perspectives, we intend to apply this classification to the entire S-PLUS iDR3, finding LSBGs at different distances and environments. We will iteratively add the new LSBGs found to the training samples to increase and improve our classification and find even more objects.

We note that this work is extendable to any multi-band survey, such as J-PLUS in the Northern hemisphere or even the Vera Rubin Telescope, which may provide an unprecedented map of LSBGs given to its high depth and sky coverage.

References

C. E. Barbosa, D. Zaritsky, R. Donnerstein, H. Zhang, A. Dey, C. Mendes de Oliveira, L. Sampedro, A. Molino, M. V. Costa-Duarte, P. Coelho, A. Cortesi, F. R. Herpich, J. A. Hernandez-Jimenez, T. Santos-Silva, E. Pereira, A. Werle, R. A. Overzier, R. Cid Fernandes, A. V. Smith Castelli, T. Ribeiro, W. Schoenell, and A. Kanaan. One Hundred SMUDGes in S-PLUS: Ultra-diffuse Galaxies Flourish in the Field. , 247(2): 46, April 2020. doi: 10.3847/1538-4365/ab7660.

Michael A. Beasley, Izaskun San Roman, Carme Gallart, Ata Sarajedini, and Antonio Aparicio. Evidence for temporal evolution in the M33 disc as traced by its star clusters. , 451(4):3400–3418, August 2015. doi: 10.1093/mnras/stv943.

Johnny P. Greco, Jenny E. Greene, Michael A. Strauss, Lauren A. Macarthur, Xzavier Flowers, Andy D. Goulding, Song Huang, Ji Hoon Kim, Yutaka Komiyama, Alexie Leauthaud, Lukas Leisman, Robert H. Lupton, Cristóbal Sifón, and Shiang-Yu Wang. Illuminating Low Surface Brightness Galaxies with the Hyper Suprime-Cam Survey. , 857(2):104, April 2018. doi: 10.3847/1538-4357/aab842.

G. Martin, S. Kaviraj, C. Laigle, J. E. G. Devriendt, R. A. Jackson, S. Peirani, Y. Dubois, C. Pichon, and A. Slyz. The formation and evolution of low-surface-brightness galaxies. , 485(1):796–818, May 2019. doi: 10.1093/mnras/stz356.

D. Tanoglidis, A. Ćiprijanović, and A. Drlica-Wagner. DeepShadows: Separating low surface brightness galaxies from artifacts using deep learning. *Astronomy and Computing*, 35:100469, April 2021. doi: 10.1016/j.ascom.2021.100469.

Aku Venhola, Reynier Peletier, Eija Laurikainen, Heikki Salo, Thorsten Lisker, Enrichetta Iodice, Massimo Capaccioli, Gijs Verdois Kleijn, Edwin Valentijn, Steffen Mieske, Michael Hilker, Carolin Wittmann, Glenn van de Ven, Aniello Grado, Marilena Spavone, Michele Cantiello, Nicola Napolitano, Maurizio Paolillo, and Jesús Falcón-Barroso. The Fornax Deep Survey with VST. III. Low surface brightness dwarfs and ultra diffuse galaxies in the center of the Fornax cluster. , 608:A142, December 2017. doi: 10.1051/0004-6361/201730696.

Dennis Zaritsky, Richard Donnerstein, Arjun Dey, Jennifer Kadowaki, Huanian Zhang, Ananthan Karunakaran, David Martínez-Delgado, Mubdi Rahman, and Kristine Spekkens. Systematically Measuring Ultra-diffuse Galaxies (SMUDGes). I. Survey Description and First Results in the Coma Galaxy Cluster and Environs. , 240(1):1, January 2019. doi: 10.3847/1538-4365/aaefe9.