Highlights of Spanish Astrophysics XII, Proceedings of the XVI Scientific Meeting of the Spanish Astronomical Society held on July 15 - 19, 2024, in Granada, Spain. M. Manteiga, F. González Galindo, A. Labiano Ortega, M. Martínez González, N. Rea, M. Romero Gómez, A. Ulla Miguel, G. Yepes, C. Rodríguez López, A. Gómez García and C. Dafonte (eds.), 2025

Unveiling features concealed in the background through Machine Learning

Suelves, L.E.[1,2], **Pearson, W.J.**[2], **Pollo, Agnieszka**[2,3]

[1] Tartu Observatory, University of Tartu, Observatooriumi 1, Tõravere 61602, Estonia
[2] National Centre for Nuclear Research, Pasteura 7, 02-093 Warszawa, Poland
[3] Astronomical Observatory of the Jagiellonian University, Orla 171, 30-001 Cracow, Poland

Abstract

Galaxy merger identification in large-scale surveys is one of the main fields benefitting from the fast embracement of Machine Learning (ML) classification models in astronomy. The low surface-brightness features presented in mergers with high resolution are easy to spot by both visual inspection and automatized ML-based methods. However, both also become more inaccurate for large surveys where the observed sources show a wide and worse variety of resolution and noise properties. In this talk, I will focus on the combination of ML, clustering, and dimensionality reduction techniques with the data reduction of astronomical images and its resulting photometric measurements. An initial Neural Network on SDSS photometry led us to find how the sky background error traces low S/N merging features. A follow-up analysis of the sky background, in the HSC North Ecliptic Pole deep field, has provided an initial understanding and quantification of the potential use of low S/N features for classification. With this work, I want to stress the benefits of interpreting the results of ML models, in contrast with the traditional black-box treatment.

1 Introduction

The evolution of galaxies and of their large scale distribution is driven by the hierarchical growth of structure. Within the current cosmological paradigm, the ΛCDM model, galaxies form within haloes of dark matter that serve as gravitational hosts. This dark matter haloes approach and merge through gravitational attraction [7, 8], making the galaxies to also merge. When galaxies interact, the gravitational tidal forces arising highly distort their morphology [10]. Therefore, galaxy mergers can trace the Universe's structure and how it has changed across cosmic time.

Galaxy merger identification is nonetheless not an easy task. The visual distortions that characterize merging interactions are very distinctive in resolved images, inspecting galaxy by galaxy is not doable for the current large number of galaxies. The shape distortions can

be parametrized through morphological measurements with short computation times [3, e.g.], or machine learning models can be trained [1, e.g.]. Moreover, pairs of galaxies in the sky can be identified to be in the process of merging by their redshift [4, e.g.].

In the oral contribution that this proceedings accompanies, I presented how we attempted to identify galaxy mergers through a Neural Network (NN). The main result was published in [9]. This proceeding reviews the data and methodology in Sect. 2, the results in Sect. 3, and the extension to more galaxies outside of the training dataset in 4.

2 Data and Methodology

The training dataset used in [9] was based on galaxy mergers from the Sloan Digital Sky Survey Data Release 6 ([2, SDSS DR6]) and teh citizen science classification of Galaxy Zoo Data Release 1 [6, GZ DR1]. The mergers were selected from [5] and the non-mergers were matched as described in [9]. The catalogue consisted of 2 680 mergers (and 2 680 non-mergers) for training plus validation, and 250 galaxies of each class for test. The data used were mainly flux measurements in the form of magnitudes, and the input parameters for measuring the aperture/fibre magnitudes. Among them, the most important one is the sky background error, `skyErr` from now.

The model used is a Neural Network (NN), which connect the input parameters to an output, in this case the probability of the galaxy to belong to the merger or non-merger labels. The NN has layer that connect subsequently between input and output, and these layers contain parameters that update each other during training. We quantified the quality of the training by the accuracy, defined as the number of correctly classified galaxies divided by the total number. More details on the NN specifications and on how the training was performed can be found in `suelves23`.

3 Results

We obtained NN training accuracies for multiple combinations of photometric parameters: magnitudes for the five bands, the ten resulting colours, and the five magnitude errors. We also considered multiple types of magnitudes measurements, being the model and fibre magnitudes the most relevant ones for this text. The model magnitude comes from the surface brightness profile fit to the galaxy, and the fibre magnitude is an aperture magnitudes with 3 arcseconds of diameter. The model errors provided a $59.06 \pm 0.76\%$ accuracy, and the fibre errors an $83.76 \pm 0.32\%$. This led us to study the fibre error and the parameters used for its calculation.

Out of the parameters that combine into the fibre errors, we found the sky background error (`skyErr`) to provide the highest accuracy. When using the logarithm of the `skyErr` in units of `counts` – analogue to ADUs –, we obtained a $92.64 \pm 0.15\%$ for training and $92.36 \pm 0.21\%$ for test. The test accuracy indicated that the NN results does not come from overfitting, i.e., the NN learning the dataset by hard instead of learning to identify it through the input. Moreover, plotting the dataset galaxies in the `skyErr` g versus r-band

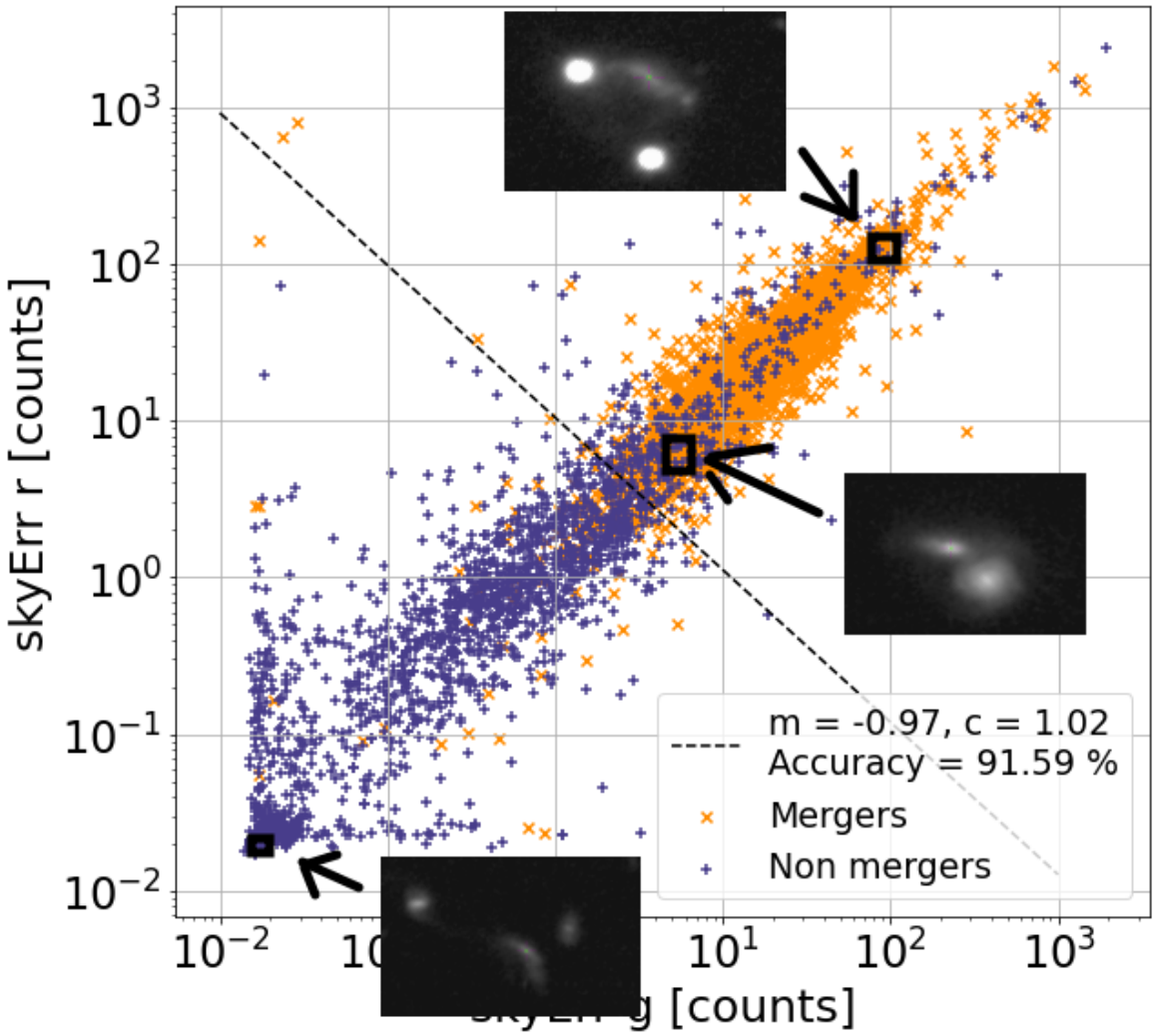

Figure 1: Training set galaxies in the 2D plane of `skyErr` in the g and r bands. The mergers are marked by orange crosses and the non-mergers by dark blue plus symbols. The boundary is the dashed black line with the parameters and classification accuracy in the label. The three cutouts correspond to galaxies located within the area of the diagram indicated by the box with which they are connected by an arrow.

plane showed that a simple decision boundary is able of providing a 91.59 % accuracy, as shown in Fig. 1. This is a very promising result: it shows how we could take advantage of a NN to learn a new property of the mergers in our training dataset, and might lead to a new promising technique to find mergers. Current work is focused on understanding better the methodology, and try to extend it to data outside the training catalogue. The following section describes part of the challenge of extending the data to more GZ DR1 galaxies, and the variation of the training galaxies depending on the magnitude.

4 Extension to other SDSS-GZ1 galaxies

The main challenges of extending the `skyErr` technique to other SDSS DR6-GZ DR1 galaxies can be observed in Fig. 1, which shows how galaxy merger cutouts around the diagram. The bottom left corner is where the majority of mergers mistakenly identified as non-mergers are located. To our knowledge, the given galaxy has low `skyErr` because its boundaries are fairly sharp. This is in contrast with the galaxy in the central area of the diagram, which is more diffused and thus shows a shape that blurs more through the background. Finally, the top right cutout, which should correspond to galaxies confidently identified as mergers, is contaminated by stars next to it. By observing many other cutouts, we found a clear majority of galaxies contaminated by nearby stars in the top-right region, including both mergers and non-mergers. The clear `skyErr` variation depending on the depth of the sources within the cutouts, led us to investigate the distribution of galaxies depending on their magnitude.

Figures 2, 3, and 4 indicate the location of galaxies with different r-band magnitudes, through a hex-bin-like plot made on the `skyErr` boundary. This is a type of plot where the area of the graph is segmented into hexagons, so that each hexagon depicts information of the sources within it through a colour code. The three plots have different magnitude intervals, above than 14.5 for Fig. 2, in (14.5, 16] for Fig. 3, and below 16 for Fig. 4.

The density panels (*right*) show how galaxies are located along the diagram mainly in two regions. One is the central area, which can be seen clearly in the two brightest cases, and the other is where the minimum `skyErr` values are, visible in the two dimmest plots in the bottom-left corner. Apparently, the brighter galaxies are in the central part of the diagram. Nonetheless, the number of galaxies per magnitude bin is quite heterogeneous, 161 for the bright galaxies shown in Fig. 2, 1 248 for the intermediate case in Fig. 3, and 4 449 for Fig. 4. In fact, the central area is overall denser for all magnitude bins, and the *central* panels of Figs. 3 and 4 indicate a large variability of r magnitude in that region. The more dim sources are located in more external regions and in the bottom-left corner, and this is clear in the density panel of Fig. 4. Besides, the standard deviation in that corner is lower than in the central regions, showing this bottom-left corner is more homogeneous even though it is more densely populated.

Inspecting the galaxy in the bottom-left corner of Fig. 1 can shed some light on why the bottom-left corner, with low sky error, is highly populated by dim sources. If the low S/N surrounding the galaxies affect the background and increase `skyErr`, galaxies with sharp boundaries will have a reduced effect. Dim galaxies tend to have more steep contours, because their LSB surroundings are barely above the background level. This is consistent with the physical interpretation of why the sky background error is able to trace mergers, and implies that there is an image depth limit for the effectiveness of a `skyErr`-based identification method. An analysis of the truncation of the boundaries of these galaxies is planned in the future in order to quantify the `skyErr` effect differently, which would help to understand better the methodology. Finally, the presence of dim galaxies in the top-right corner can also be understood through Fig. 1: those dim galaxies are quite likely to be contaminated by nearby stars.

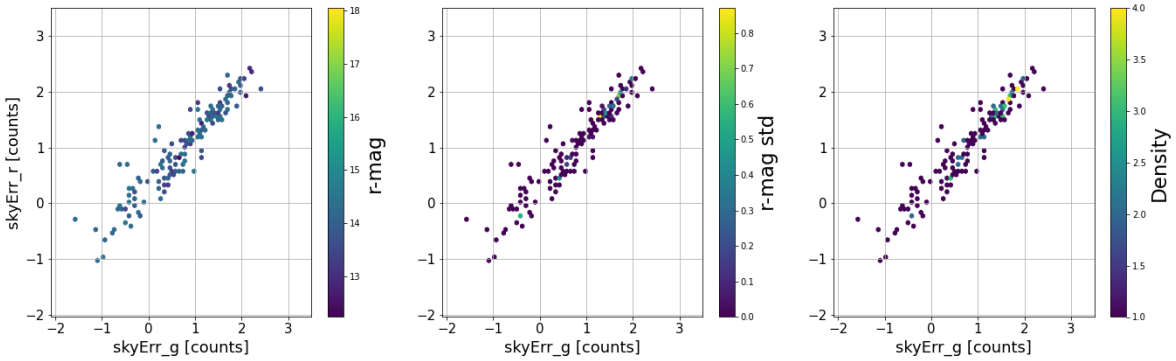

Figure 2: Three hexagonal bin plots, that show the median r-band magnitudes (*left*), standard deviation (*centre*) and number density (*right*) of the galaxies located in the area covered by the hexagons. The axes are the g and r-band sky background errors, analogously to Fig. **??** showing the training galaxies. In this case, the galaxies are those with r-band magnitudes above 14.5. The hexagon colour-map is indicated on the right of each panel, together with the variable that said colour-map indicates.

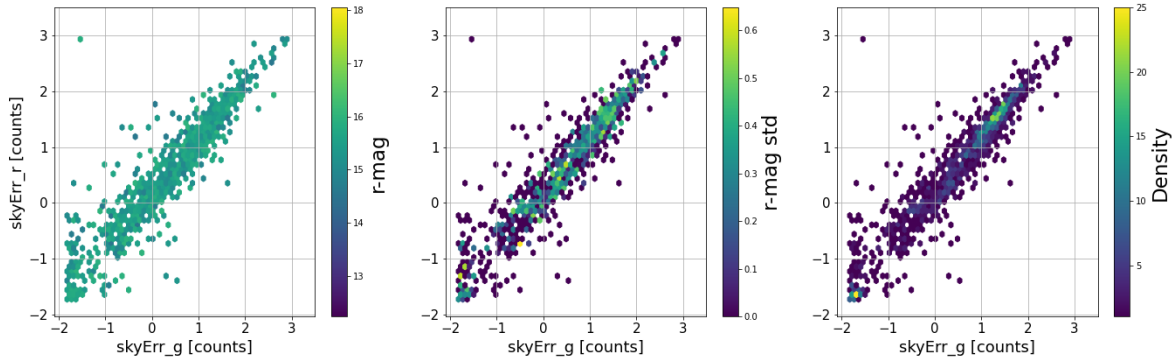

Figure 3: Analogous to Fig. 2, but for training set galaxies in the r-band magnitude interval (14.5, 16].

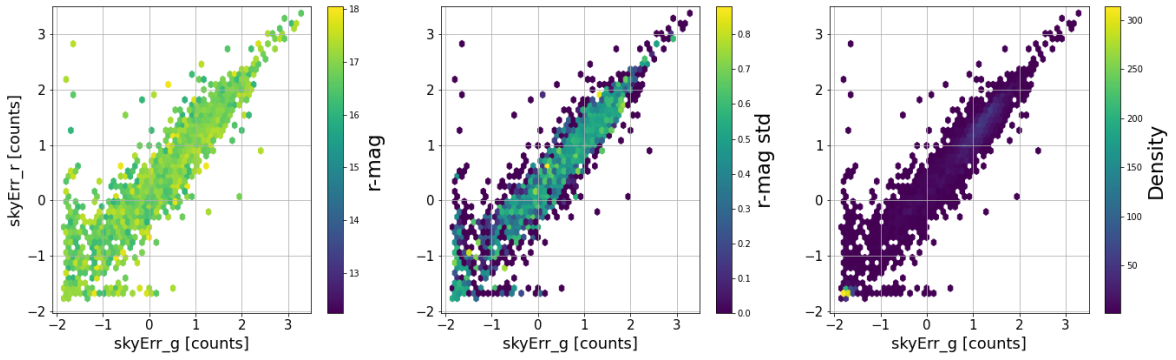

Figure 4: Analogous to Fig. 2, but for training set galaxies with r-band magnitude bellow 16.

Acknowledgments

L.E. Suelves was supported by the Estonian Ministry of Education and Research (grant TK202), Estonian Research Council grant (PRG1006), and the European Union's Horizon Europe research and innovation programme (EXCOSM, grant No. 101159513). W.J. Pearson has been supported by the Polish National Science Center projects (NCN) UMO-2020/37/B/ST9/00466 and UMO-2023/51/D/ST9/00147. A.Pollo has been supported by NCN UMO 2023/50/A/ST9/00579.

References

[1] Ackermann, S., Schawinski, K., Zhang, C., Weigel, et al., 2018, ApJ, 479(1), 415–425

[2] Adelman-McCarthy, J.K., Ag¨ueros, M.A., Allam, S.S., et al., 2008, ApJ, 175(2), 297–313

[3] Conselice, C.J., Bershady, M.A., Jangren, A., 2000, ApJ, vol. 529(2), 886–910

[4] Duncan, K., Conselice, C.J., Mundy, C.e.a., 2019, ,ApJ 876(2), 110

[5] Darg, D.W., Kaviraj, S., Lintott, C.J., et al., 2010, MNRAS, 401(2), 1043–1056

[6] Lintott C., Schawinski K., Bamford S, et al., 2011, MNRAS, 410(1), 166–178

[7] White, S.D.M., Frenk, C.S., 1991, ApJ, 379, 52

[8] White, S.D.M., Rees, M.J., 1978, MNRAS , 183, 341–358

[9] Suelves, L.E., Pearson, W.J., Pollo, A., 2023, A&A, 669, A141

[10] Toomre, A., Toomre, J., 1972, ApJ, 178, 623–666