A feedback-loop approach for galaxy physical properties estimation

Davide Zago[1*], Giovanni Bonetta[2], Rossella Cancelliere[1], Mario Gai[3]

1- University of Turin - Department of Computer Science
via Pessinetto 12 - 10149 Turin, Italy
2- Fondazione Bruno Kessler, via Sommarive, 18 - Povo 38123 Trento, Italy
3- Istituto Nazionale di Astrofisica, Osservatorio Astrofisico di Torino
V. Osservatorio 20, 10025 Pino Torinese (TO), Italy

Abstract.
Ongoing and forthcoming surveys promise great advances in our understanding of the Universe content and history, thanks to unprecedented improvements in the size and precision of observation datasets. On a cosmological scale, galaxies characteristics may be summarised by three main features, namely their redshift, stellar content mass and star formation rate, evolving throughout their lifetime. They are usually estimated from a set of photometric measurements, mapping their spectral emission. In this context, we propose a machine learning approach where we first evaluate redshift from the photometric data, and then merge it with them through a feedback loop, for subsequent estimation of the three desired parameters. In spite of its simplicity, our approach matches the performance of, and in some cases outperforms, significantly more complex previous tools exploiting also images. It achieves correct estimates on the near totality of instances for redshift and stellar mass, decreasing to about 70% on the more difficult case of SFR estimation.

1 Introduction

Understanding galaxy evolution and its governing mechanisms is central to astrophysics, supported by extensive multi-wavelength datasets from surveys and missions. The European Space Agency's Euclid mission (`www.euclid-ec.org`, [1, 2]) exemplifies this progress, offering visible and near-IR data for ~ 1.5 billion objects and NIR spectra for > 35 million sources, covering $\sim 1/3$ of the sky. Key properties of galaxy evolution include redshift, stellar mass, and star formation rate (SFR). These are traditionally derived via spectral energy distribution (SED) fitting, which compares observed data to theoretical models [3, 4]. However, scaling this approach to massive datasets requires substantial computational resources and expertise. Machine learning (ML) offers a powerful alternative, enabling efficient and accurate estimation of physical properties, as evidenced by the increasing adoption of such approaches within the astrophysical community. This study focuses on ML-based estimation of redshift, stellar mass, and SFR using photometric data within the Euclid mission's framework: the scientific objectives and overarching framework are defined in [5]. Our work utilizes

*Corresponding author. Email: davide.zago@unito.it

the same dataset of real photometric data, with the permission of that paper's authors. In [6], we proposed a unified neural network for simultaneous redshift, stellar mass, and SFR estimation, rather than implementing three dedicated models, as done in [5], leveraging interdependencies among these properties. Expanding on this, we introduce a novel approach where the network estimates redshift and uses it as an input for subsequent estimation of the two remaining desired parameters. This feedback mechanism seems to capture underlying correlations, improving comprehensive parameter estimation. The obtained results confirm that the image input components used in [5] and [6] are no more necessary and good performance can be reached with a lighter, less computational expensive architecture.

2 Dataset and Metrics

The data set used in our work was derived from the COSMOS2015 multi-wavelength public catalogue [7]. The data is described in detail in [5], in terms of dataset structure and characteristics, and is used without modifications in our work. The custom catalogue is inspired to the Euclid Wide Survey [8], and includes in a tabular form the four Euclid filters, i.e. I_E, Y_E, J_E, H_E with the addition of the u band from the Canada-France Imaging Survey (CFIS), and of the Sloan Digital Sky Survey (SDSS) magnitudes g, r, i and z. We use all nine photometric data, i.e. the four Euclid ones plus the five ground-based ones, as in [5], as inputs to our system in order to diagnose redshift, stellar mass and SFR.

The Mean square error (MSE) is used as training loss function and to evaluate the quality of the model. Several metrics are also used to measure the performance, following [5] and references therein, i.e. the fraction of outliers (f_{out}), the bias and the Normalised Median Absolute Deviation (NMAD). The fraction of outliers f_{out} corresponds to the amount of over- or under-estimated data, relative to the size of the dataset. The bias and NMAD are used to provide an indication of the statistical distribution of the estimates. In particular, it is expected that this distribution is approximated by a Gaussian, so that Δz indicates the redshift mean discrepancy (ideally zero), while the NMAD is related to its standard deviation. In the following, the subscript 'in' refers to the target values, and the subscript 'out' refers to the estimated values.

Redshift (z). As in [5], a prediction is called an outlier if $\frac{|z_{\text{out}}-z_{\text{in}}|}{1+z_{\text{in}}} > 0.15$. Bias Δz and NMAD are defined in eq.(1).

$$\Delta z = \text{median}\left[\frac{z_{\text{out}} - z_{\text{in}}}{1 + z_{\text{in}}}\right], NMAD = 1.48\ \text{median}\left[\frac{|z_{\text{out}} - z_{\text{in}}|}{1 + z_{\text{in}}}\right] \quad (1)$$

Stellar mass (M_*). As in [5], we estimate for the entire sample the fraction of outliers, defined as galaxies for which the stellar mass is overestimated or underestimated by a factor two (~ 0.3 dex): a mass prediction is considered an outlier if $\left|\log_{10}\left(\frac{M_{*,\text{out}}}{M_{*,\text{in}}}\right)\right| > 0.3$. Bias ($\Delta M_*$) and NMAD are defined in eq.(2).

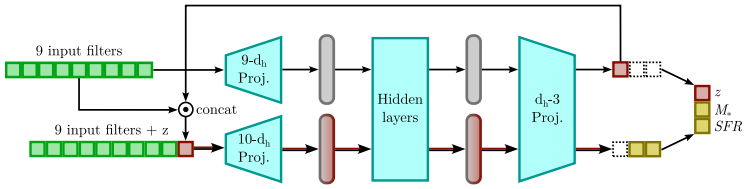


Fig. 1: Diagram of the model architecture. We interleave linear transformations with ReLU activation functions and dropout [9].

$$\Delta M_* = \text{median}\left[\log_{10}\left(\frac{M_{*,\text{out}}}{M_{*,\text{in}}}\right)\right], NMAD = 1.48\ \text{median}\left[\left|\log_{10}\left(\frac{M_{*,\text{out}}}{M_{*,\text{in}}}\right)\right|\right] \quad (2)$$

Star Formation Rate (SFR). Similarly to the stellar mass case, we define outlier the estimation with SFR incorrect by at least a factor two, i.e. $\left|\log_{10}\left(\frac{SFR_{\text{out}}}{SFR_{\text{in}}}\right)\right| > 0.3$. Bias ($\Delta SFR$) and NMAD are defined in eq.(3):

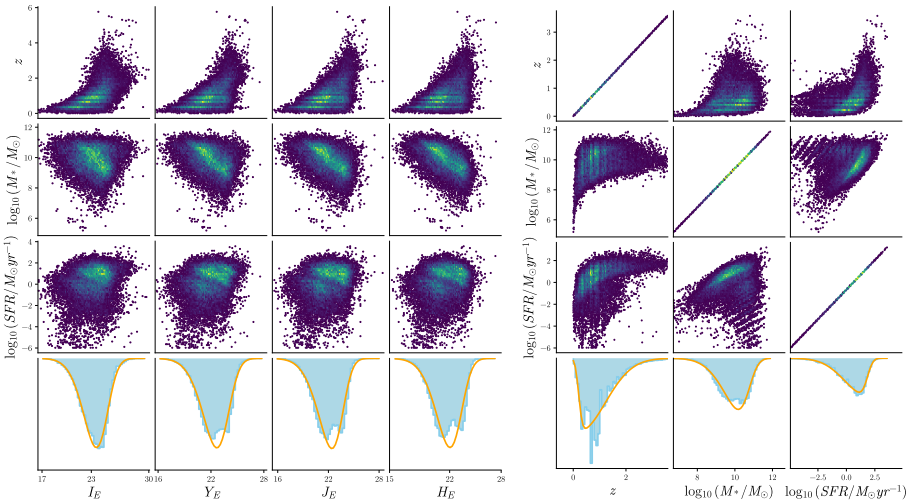
$$\Delta SFR = \text{median}\left[\log_{10}\left(\frac{SFR_{\text{out}}}{SFR_{\text{in}}}\right)\right], NMAD = 1.48\text{median}\left[\left|\log_{10}\left(\frac{SFR_{\text{out}}}{SFR_{\text{in}}}\right)\right|\right] \quad (3)$$

3 Experiments and Results

In order to estimate redshift, stellar mass and SFR, in our experiments we use a multilayer perceptron, named Z-MLP, enhanced with a feedback loop (see fig.1), in order to capture the existing correlations evidenced in fig.2. In the first feedforward phase, the nine photometric bands described in sec.2 are projected to the latent dimension d_h, and then refined through a series of hidden layers (2 our case). In the obtained triple y_1, we collect the first element as predicted redshift, and we project its concatenation with the 9 input filters to the hidden dimension. The new representation is processed through the same layers and results in the new triple y_2, from which we identify stellar mass and star formation rate estimations in the second and third elements.

The dataset is split into training, validation and test sets, respectively sized to 90%, 5% and 5% of the total, so that 24,605 instances are used for training, 1,367 for validation and 1,367 for test. Data normalization and removal of anomalous data are performed. The training loss is the sum of the MSEs of the three target predictions. In all experiments, a batch of 1024 data instances is used. For each experiment, the training is carried for 3000 epochs. The validation loss is tracked, and at the end of training the model with the best overall validation performance is saved.

This system is trained using five different seeds. In this experiment we consider the best run, i.e. the run achieving the lowest value of the validation loss function (hereafter identified with subscript *bst*), and the average of the five runs (identified with subscript *avg*), which provides an indication of the spread which



(a) Correlation between the 4 Euclid filter magnitudes and galaxy properties.

(b) Correlation between galaxy properties.

Fig. 2: Correlation in the COSMOS-DASH dataset. Scatter plots display the correlation of a given measure (input filter or target value) to a galaxy property. Colors towards yellow represent more densely populated areas. Last row of fig. 2a, 2b display the data distribution with a skewed normal approximation over a 50-bins histogram.

might be expected on the results. The experiments are performed on a Linux-5.13.0 computer equipped with an Intel Core i7 8th Gen processor, 8 CPUs and a NVIDIA TITAN RTX GPU (24gb). The software is written in Python 3.10.15 with Pytorch. The time required to train the network is about three hours. The code will be shared after publication.

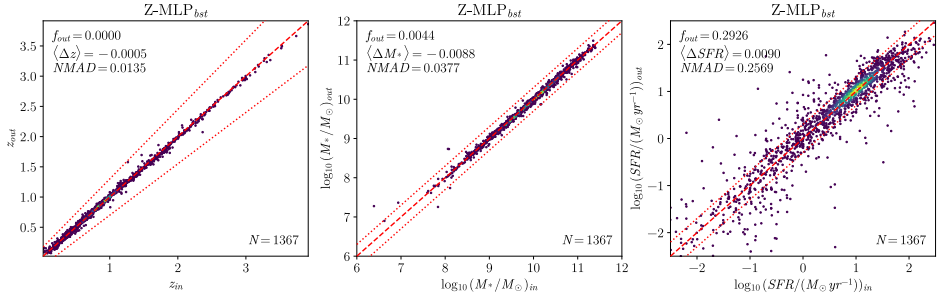
We compare outputs from different models in terms of outlier fraction, bias and NMAD as described in Section 2. For the sake of a better understanding we report i) the results from [5] for the SED fitting model and CNN, where CNN is a multi-modal system that exploits also images and ii) the result from [6] for MLP and Fusion Network (FN), which exploits images too.

Redshift. The main features of our redshift estimation are reported in Table 1 in statistical terms, and illustrated in Fig. 3a. All ML models achieved significantly better results than the reference SED fitting method, and in particular for Z-MLP the f_{out} is three orders of magnitude smaller. MLP is robust in estimating the redshift, with f_{out}, bias and NMAD in line with FN. The best model turns out to be Z-MLP, featuring no outlier as highlighted in Fig. 3a, where the predicted redshift values are always within the Δz cone boundaries. Besides, Δz results are the lowest ones.

Stellar mass. The main features of stellar mass estimates are reported in Table

Model	Redshift (z)			Stellar Mass (M_*)			Star Formation Rate (SFR)		
	$f_{\rm out}$	Δ	NMAD	$f_{\rm out}$	Δ	NMAD	$f_{\rm out}$	Δ	NMAD
SED	0.127	-0.002	0.045	0.135	0.002	0.12	0.62	-0.06	0.64
$\rm CNN_{avg}$	0.003	-0.001	0.021	0.010	**0.001**	0.04	0.45	-0.06	0.39
$\rm FN_{avg}$	0.001	-0.002	0.009	0.005	**-0.001**	0.04	0.30	**0.001**	**0.24**
$\rm MLP_{avg}$	0.002	-0.001	**0.008**	0.006	0.007	**0.03**	0.34	-0.01	0.27
$\rm Z\text{-}MLP_{avg}$	0.0007	**-0.0004**	0.010	0.007	-0.008	**0.03**	0.30	-0.006	0.26
$\rm CNN_{bst}$	0.002	0.005	0.028	0.011	0.006	0.05	0.44	0.02	0.38
$\rm FN_{bst}$	0.0007	-0.002	0.011	**0.004**	-0.01	0.05	0.30	0.02	0.25
$\rm MLP_{bst}$	0.002	-0.001	0.011	0.006	0.005	0.05	0.34	-0.02	0.27
$\rm Z\text{-}MLP_{bst}$	**0.0000**	-0.0005	0.013	**0.004**	-0.009	0.04	**0.29**	0.009	0.26

Table 1: Results for redshift, stellar mass and SFR estimation.



(a) Redshift (b) Stellar mass (c) Star formation rate

Fig. 3: Z-MLP diagnostics performance on redshift (left), stellar mass (middle) and SFR (right).

1 and illustrated in Fig. 3b. The SED model provides the worst results among the tested cases. Z-MLP has slightly better performance than MLP and similar performance with FN on outliers $f_{\rm out}$, with a somewhat worse bias ΔM_* and comparable, or slightly better, NMAD.

SFR. The results for the SFR estimation are reported in Table 1 and in Fig 3c. Our method leads to significant improvements on outliers ($f_{\rm out}$) over the SED fitting method (62%), and performs better than FN and MLP models. Also, the bias and NMAD are significantly smaller than that from other techniques, with the exception of FN.

Reduced compute time is the major advantage of the proposed Z-MLP: training time for one epoch require less than 4 seconds, while 3 minutes for the FN. Total inference time on the test set is around one second for Z-MLP and 17 seconds for FN. These data highlight a time saving of 98% in training, and 94% in inference. The notable improvements achieved with our Z-MLP solution highlight the effectiveness of a straightforward neural network approach that:

- uses only photometric features to estimate the properties of target galaxies, so representing a much faster inference method.

- incorporates a feedback loop that integrates the redshift prediction into the estimation of the other correlated parameters.

4 Conclusions

The result improvements achieved appear to be related to two different aspects, i.e. the simultaneous estimation of the three physical properties and the use of the feedback loop, which may be expected to bear connections in classes of objects, giving our tool a chance to learn the "shape" underlying our data distribution. In general, Z-MLP achieves best results with respect to the simple, similar architecture MLP and comparable results w.r.t. FN, but needing extremely less training and inference processing times. It provides competitive performance with significantly reduced complexity, since it only works on photometric data without exploiting galaxy images, whose processing is computationally demanding. It appears as an interesting tool which may be conveniently retained for comparison and verification purposes. The SFR estimation still remains much more noisy than redshift and stellar mass, a characteristic possibly implicit in the natural spread of SFR throughout the dataset.

Acknowledgements We acknowledge usage of the dataset from [5], kindly provided by that paper's Authors. MG acknowledges support by the Agenzia Spaziale Italiana (ASI) through contract 2018-24-HH.0 and its Addendum 2018-24-HH.1-2022.

References

[1] René Laureijs et al. Euclid: ESA's mission to map the geometry of the dark universe. In *Space Telescopes and Instrumentation 2012: Optical, Infrared, and Millimeter Wave*, volume 8442 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, September 2012.

[2] Yannick Mellier et al. Euclid. I. Overview of the Euclid mission. *arXiv e-prints*, page arXiv:2405.13491, May 2024.

[3] L. Bisigello et al. Recovering the Properties of High-redshift Galaxies with Different JWST Broadband Filters. *ApJS*, 231(3):1–17, July 2017.

[4] K. Iyer, E. Gawiser, R. Davé, P. Davis, and et al. The SFR-M_* Correlation Extends to Low Mass at High Redshift. *ApJ*, 866(2):120, October 2018.

[5] Euclid Collaboration: L. Bisigello et al. Euclid preparation - XXIII. Derivation of galaxy physical properties with deep machine learning using mock fluxes and H-band images. *Monthly Notices of the Royal Astronomical Society*, 520(3):3529–3548, April 2023.

[6] Mario Gai et al. Simultaneous derivation of galaxy physical properties with multimodal deep learning. *MNRAS*, 532(2):1391–1401, 2024.

[7] C. Laigle, H. J. McCracken, O. Ilbert, B. C. Hsieh, and et al. The COSMOS2015 Catalog: Exploring the $1 < z < 6$ Universe with Half a Million Galaxies. *ApJS*, 224(2):1–24, June 2016.

[8] Euclid Collaboration: Scaramella, R., Amiaux, J., Mellier, Y., Burigana, C., and et al. Euclid preparation - i. the euclid wide survey. *A&A*, 662:A112, 2022.

[9] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.