The search for galaxy cluster members with deep learning of panchromatic HST imaging and extensive spectroscopy*

G. Angora1,3, P. Rosati1,2,9, M. Brescia3, A. Mercurio3, C. Grillo4,5,6, G. Caminha7, M. Meneghetti2, M. Nonino8, E. Vanzella2, P. Bergamini2, A. Biviano8, and M. Lombardi4

1 Department of Physics and Earth Science of the University of Ferrara, Via Saragat 1, 44122 Ferrara, Italy
e-mail: gius.angora@gmail.com
2 INAF – Astronomical Observatory of Bologna, via Gobetti 93/3, 40129 Bologna, Italy
3 INAF – Astronomical Observatory of Capodimonte, Salita Moiariello 16, 80131 Napoli, Italy
4 Department of Physics of the University of Milano, via Celoria 16, 20133 Milano, Italy
5 Dark Cosmology Centre, Niels Bohr Institute, University of Copenhagen, Lyngbyvej 2, 2100 Copenhagen, Denmark
6 INAF – IASF Milano, Via A. Corti 12, 20133 Milano, Italy
7 Kapteyn Astronomical Institute, University of Groningen, Postbus 800, 9700 AV Groningen, The Netherlands
8 INAF – Astronomical Observatory of Trieste, via G. B. Tiepolo 11, 34131 Trieste, Italy
9 INFN, Sezione di Ferrara, Via Saragat 1, 44122 Ferrara, Italy

Received 1 August 2020 / Accepted 16 September 2020

ABSTRACT

Context. The next generation of extensive and data-intensive surveys are bound to produce a vast amount of data, which can be efficiently dealt with using machine-learning and deep-learning methods to explore possible correlations within the multi-dimensional parameter space.

Aims. We explore the classification capabilities of convolution neural networks (CNNs) to identify galaxy cluster members (CLMs) by using Hubble Space Telescope (HST) images of fifteen galaxy clusters at redshift 0.19 ≤ z ≤ 0.60, observed as part of the CLASH and Hubble Frontier Field programmes.

Methods. We used extensive spectroscopic information, based on the CLASH-VLT VIMOS programme combined with MUSE observations, to define the knowledge base. We performed various tests to quantify how well CNNs can identify cluster members on the basis of imaging information only. Furthermore, we investigated the CNN capability to predict source memberships outside the training coverage, in particular, by identifying CLMs at the faint end of the magnitude distributions.

Results. We find that the CNNs achieve a purity-completeness rate ≳90%, demonstrating stable behaviour across the luminosity and colour of cluster galaxies, along with a remarkable generalisation capability with respect to cluster redshifts. We concluded that if extensive spectroscopic information is available as a training base, the proposed approach is a valid alternative to catalogue-based methods because it has the advantage of avoiding photometric measurements, which are particularly challenging and time-consuming in crowded cluster cores. As a byproduct, we identified 372 photometric cluster members, with mag(F814) < 25, to complete the sample of 812 spectroscopic members in four galaxy clusters RX J2248-4431, MACS J0416-2403, MACS J1206-0847 and MACS J1149+2223.

Conclusions. When this technique is applied to the data that are expected to become available from forthcoming surveys, it will be an efficient tool for a variety of studies requiring CLM selection, such as galaxy number densities, luminosity functions, and lensing mass reconstruction.

Key words. Galaxy: general – galaxies: photometry – galaxies: distances and redshifts – techniques: image processing – methods: data analysis

1. Introduction

Over the past decade, the field of astrophysics has been experiencing a true paradigmatic shift, moving rapidly from relatively small data sets to the big data regime. Dedicated survey telescopes, both ground-based and space-borne, are set to routinely produce tens of terabytes of data of unprecedented quantity and complexity on a daily basis. These volumes of data can be dealt with through a novel framework, delegating most of the work to automatic tools and by exploiting all advances in high-performance computing, machine learning, data science and visualisation (Brescia et al. 2018). The paradigms of machine learning (ML) and deep learning (DL) paradigms embed the intrinsic data-driven learning capability to explore huge amounts of multi-dimensional data by searching for hidden correlations within the data parameter space.

Here, we explore the application of ML techniques in the context of studies of galaxy clusters, more specifically, to identify cluster members (CLMs) based on imaging data alone. In fact, obtaining a highly complete sample of spectroscopic members is an extremely expensive and time-consuming task, which can be simplified and accelerated thanks to the use of a limited amount of spectroscopic information in training ML methods.

Disentangling CLMs from background and foreground sources is an essential step in the measurement of physical
properties of galaxy clusters, measuring, for example, the galaxy luminosity and stellar mass functions (e.g. Annunziatella et al. 2016, 2017), in addition to studies of the cluster mass distribution via strong and weak lensing techniques (e.g. Caminha et al. 2017a, 2019; Lagattuta et al. 2017; Medezinski et al. 2016). In particular, the study of the inner mass substructure of cluster cores with high-precision strong-lensing models and their comparison with cosmological simulations requires the simultaneous identification of background multiply lensed images and member galaxies to separate the sub-halo population from the cluster projected total mass distribution (e.g. Grillo et al. 2015; Bergamini et al. 2019). Such studies provide tests for structure-formation models and the cold dark matter paradigm (Diemand & Moore 2011; Meneghetti et al. 2020). The need for efficient and reliable methods to identify cluster member galaxies from the overwhelming population of foreground and background galaxies will become particularly pressing when a vast amount of photometric information becomes available with forthcoming surveys with, for example, the Large Synoptic Survey Telescope (LSST, Ivezić et al. 2019) and Euclid (Laureijs et al. 2014).

Owing to their ability to extract information from images, convolutional neural networks (CNNs, LeCun et al. 1989) have been widely used in several astrophysical applications, generally showing higher robustness and efficiency with respect to traditional statistical approaches. For example, they have been applied to phase-space studies of mock distributions of line-of-sight velocities of member galaxies at different projected radial distances. These DL techniques were able to reduce the scatter of the relation between cluster mass and cluster velocity dispersion by ~35\% and by ~20\% when compared to similar ML methods, for instance, the support distribution machines (Ho et al. 2019). A similar DL approach has been successfully used to predict cluster masses from mock Chandra X-ray images, by limiting the parameter space to photometric features only, thus minimising both bias (−5\%) and scatter (−12\%), on average (Ntampaka et al. 2015, 2016, 2019). Such CNNs were also successfully used to discriminate between degenerate cosmologies, including modified gravity and massive neutrinos, by inspecting simulated cluster mass maps. Merten et al. (2019) showed that the DL techniques are able to capture distinctive features in maps mimicking lensing observables, improving the classification success rate with respect to classical estimators and map descriptors.

In recent years, the selection of CLMs has been addressed in several ways: via the classical identification of the members’ red-sequence in colour-magnitude diagrams, aided by spectroscopic measurements (e.g. Caminha et al. 2019 for strong lensing applications); by measuring photometric redshifts with a ML approach based on the so-called multi-layer perceptron (Postman et al. 2012), that is, the optical and NIR bands, excluding the UV filters for which the signal-to-noise ratio (S/N) is too low. Due to differences in velocity.

Cluster images were acquired by the HST ACS and WFC3 cameras as part of the CLASH (Postman et al. 2012) and HFF (Lott et al. 2017) surveys. The images were calibrated, reduced and then combined into mosaics with spatial resolutions of 0.065″ (see Koekemoer et al. 2007, 2011). The fifteen clusters used in our study are shown in Fig. 1. Colour images were produced with the Trilogy code (Coe et al. 2012), by combining HST filters from the optical to the near-infrared (NIR). Among the 16 available HST filters used in our experiments, we considered bands covering the spectral range 4000 Å–16 000 Å (Postman et al. 2012), that is, the optical and NIR bands, excluding the UV filters for which the signal-to-noise ratio (S/N) of faint CLMs was too low.

For each spectroscopic source within the HST images, we extracted a squared cut-out with a side of ~4″ (64 pixels), centered on the source position. A sample of the dataset is shown in Fig. 2, where CLMs were extracted from five clusters: Abell 383 (A383, z = 0.188), RX J2248-4431 (R2248, z = 0.346), MACS J0416-2403 (M0416, z = 0.397), MACS J1206-0847 (M1206, z = 0.439), and MACS J1149+2223 (M1149, z = 0.542). Due to different pointing strategies and the fields of view of HST cameras, many sources do not have a complete photometric coverage, especially in the IR range. As a result, these objects with missing information were not useful for the training process (Batista & Monard 2003; Marlin 2008; Parker 2010). With the aim of maximising the number of training samples with available spectroscopic redshift information, we chose four different band configurations:

\[ \text{Also known as Abell S1063.} \]
G. Angora et al.: Galaxy cluster membership with deep learning

Fig. 1. Colour-composite images of the 15 clusters included in our analysis, obtained by combining HST bands from optical to near IR. The images are squared cut-outs, ~130′ across, centred on the cluster core.

3. Methodology

In this work, we discuss the results achieved by a VGGNET-like model, which is a CNN implementation inspired by the VGG network proposed by Simonyan & Zisserman (2014).

We tested different network architectures, e.g. Residual Net X (He et al. 2015; Xie et al. 2016) and Inception Net (Szegedy et al. 2014). Due to their lower performances, we limited the description of the results to the VGGNET-like model, to avoid weighing down the text.

As is customary in applications of ML methods, the data require a preparation phase, which, in this case, consisted of a data augmentation procedure, that is meant to construct a consistent labelled sample, followed by a partitioning of the dataset into training, validation, and blind testing sets.

Regarding data augmentation, given the relatively small sample of spectroscopic sources with respect to the typical size of the knowledge base required by supervised ML experiments, we increased the training set by adding images of spectroscopic sources, obtained from the original ones, through rotations and flips. The inclusion of these images in the training set also offered the possibility to make the network invariant to these operations, which works as an advantage for astronomical images as there is no defined orientation for the observed sources.

Concerning the partitioning of the data set, in order to fully cover the input parameter space, we opted for a stratified k-fold partitioning approach (Hastie et al. 2009; Kohavi 1995): the whole data set was split into k = 10 non-overlapping folds, of which, iteratively, one extracted subset was used as a blind test set, while the others were taken as a training set. Such an approach has several advantages: (i) increase of the statistical significance of the test set; (ii) the blind test is performed only on original images; and (iii) complete coverage of both training and test sets, keeping them well-separated at the same time.

The classification performance, obtained through all the experiments performed by this procedure, was evaluated by adopting a set of statistical estimators, directly derived from the classification confusion matrix (Stehman 1997), namely, the classification efficiency (AE), averaged over the two classes (members and non-members), the purity (pur), the completeness (comp), and the harmonic mean of purity and completeness (F1, see Appendix A.3). The last three estimators have been measured for each class. Completeness (also known as recall) and
purity (also known as precision) are the most interesting estimators, suitable for measuring the quality of the classification performed by any method. The completeness, in fact, measures the capability to extract a “complete” set of candidates of a given class, while purity estimates the capability of selecting a “pure” set of candidates (thus, minimising the contamination). Therefore, the classification quality is usually based on either one of such two estimators or their combination, depending on the specific interest of an experiment (D’Isanto et al. 2016). In our case, we were most interested in finding the best trade-off between both estimators for the cluster members. The statistical evaluation was completed by also using the receiver operating characteristic curve (ROC, Hanley & McNeil 1982), which is a diagram where the true positive rate (TPR, i.e. the completeness rate) is plotted versus the false positive rate (FPR, i.e. the contamination rate, which corresponds to 1-purity) by varying the membership probability threshold. The model performances are measured in terms of the area under the curve (AUC), thus providing an aggregate measure of performance across all possible classification thresholds.

A full description of the data preparation procedure and the statistical estimators is given in Appendix A, while details about the architecture and configuration of the DL model are reported in Appendix B.

4. Experiments

In this section, we describe several experiments designed to test the performance of the CNNs and other methods. Specifically, with the data described in Sect. 2, we performed the following tests or experiments:
– EXP1: efficiency of the DL approach by stacking the data of all the clusters in terms of:
  – EXP1a: global evaluation
  – EXP1b: redshift-dependence, namely separating CLMs into redshift bins;
– EXP2: magnitude or colour dependence, by stacking data of a group of three clusters and varying their redshift range through:
  – EXP2a: separating bright and faint sources
  – EXP2b: separating red and blue galaxies
– EXP3: a comparison of performances of our image-based CNN technique with other approaches, based on photometric measurements of field and cluster galaxies.

4.1. EXP1: Combination of all clusters

At the first stage, we evaluated the global efficiency of a DL approach including all the available clusters, regardless of their redshift (ranging between 0.2 and 0.6), by exploring different combinations of photometric bands (as described in Sect. 2) and assembling the data set by stacking the information from all the images extracted from our cluster sample. We wanted to verify that DL models, given their intrinsic generalisation capabilities, were able to learn how to disentangle cluster members from non-member (foreground or background) sources, independently from the cluster redshift (EXP1a). This although their members have different characteristics, such as apparent magnitudes or sizes, and also different signal-to-noise ratio at a fixed apparent magnitude, due to the different image depths. The results are shown in Fig. 4 and Table D.1, as a function of the band configuration, described in Sect. 2.

For NCLM, we found similar values of the average efficiency (87%−89%), the purity (stable around ~90%) and the F1-score (with variations within 1.5%), regardless of band configuration. On the other hand, the CLM identification was, in general, characterised by larger variation (83%−91%) in the statistical estimators. With the mixed configuration, CNN achieved the best performances for CLM and it was also very stable in terms of NCLM, reaching an overall efficiency of ~89%.
We also show, in Appendix D.1, the estimators obtained for each cluster (Table D.2 and Fig. D.1). This analysis confirmed that the mixed* combination showed the highest statistical values for all the thirteen clusters. Moreover, as expected, we demonstrated that there is a clear improvement of classification capabilities as the number of sources increases (an accuracy gain of $\sim 2.3\%$ for an increment of 500 samples). Furthermore, fluctuations of these estimators tend to be better constrained for a large set of objects, stabilising around $3\%$ when the number of samples is $\ge 2000$ and showing an average reduction of $\sim 9\%$ by quadrupling the number of sources.

Since the training set we used in this study was composed of galaxies spanning a large redshift range, as part of EXP1, we investigated whether any dependence on redshift is present. To this aim, the CLM redshift range was split into five equal-sized bins ($\sim 280$ samples). The performances and fluctuations related to the mixed* band combination are shown in Fig. 5, while details on the metrics are given in Table D.3. Despite the dissimilarities between galaxies at different depths, the CNN did not seem to be affected by the CLM redshift. In fact, CNN performances achieved in different redshift bins were all comparable, with a dispersion included within 0.04–1σ for the 65% of cross-compared estimator pairs and a mean separation of $\sim 0.8\sigma$.

Since the mixed* band combinations provided the best results, all further experiments in the next sections refer to this band configuration.

Table 1. Cluster sample description.

| Cluster           | $z_{\text{cluster}}$ | $z_{\text{min}}$ | $z_{\text{max}}$ | CLMs  | NCLMs  | CLMs  | NCLMs  | CLMs  | NCLMs  | CLMs  | NCLMs  | Refs. |
|-------------------|-----------------------|------------------|------------------|-------|--------|-------|--------|-------|--------|-------|--------|-------|
| Abell 383         | A383                  | 0.188            | 0.176            | 0.200 | 59     | 51    | 91     | 79    | 59     | 51    |       | (1, 2) |
| RX J1219+0005     | R1219                 | 0.234            | 0.222            | 0.246 | 47     | 124   | 66     | 132   | 40     | 118   |       | (3, 1) |
| Abell 2744        | A2744                 | 0.308            | 0.288            | 0.331 | 156    | 279   | 166    | 203   | 166    | 177   |       | (4, 1) |
| MS 2137-2353      | MS2137                | 0.316            | 0.303            | 0.329 | 45     | 49    | 80     | 45    | 49     |       | (3, 1) |
| RX J2248-4431     | R2248                 | 0.346            | 0.332            | 0.359 | 131    | 112   | 203    | 166   | 177    | 86    |       | (5, 1) |
| MACS J1311-0310   | M1311                 | 0.352            | 0.338            | 0.365 | 68     | 97    | 80     | 110   | 65     | 96    |       | (3, 1) |
| MACS 1115+0129    | M1115                 | 0.352            | 0.338            | 0.365 | 78     | 69    | 116    | 111   | 62     | 55    |       | (3, 1) |
| Abell 370         | A370                  | 0.375            | 0.361            | 0.389 | 224    | 311   | Only frontier-field bands |
| MACS J0416-2403   | M0416                 | 0.397            | 0.382            | 0.410 | 237    | 277   | 266    | 287   | 227    | 230   |       | (7, 8, 9, 1) |
| MACS J1206-0847   | M1206                 | 0.439            | 0.425            | 0.454 | 172    | 216   | 226    | 242   | 149    | 203   | (10, 1) |
| MACS J0329-0211   | M0329                 | 0.450            | 0.435            | 0.464 | 74     | 76    | 104    | 106   | 66     | 73    |       | (3, 1) |
| RX J1311-0310     | M1311                 | 0.494            | 0.477            | 0.507 | 52     | 54    | 69     | 95    | 52     | 54    |       | (3, 1) |
| MACS J1149+2223   | M1149                 | 0.542            | 0.527            | 0.558 | 141    | 127   | 149    | 203   | 129    | 202   |       | (11, 12, 1) |
| MACS 2137-0741    | M2137                 | 0.587            | 0.571            | 0.603 | 89     | 102   | 92     | 103   | 89     | 101   |       | (1, 3) |

Notes. The name of the clusters, their redshift and their spectroscopic range to identify CLMs are reported in the first 5 columns. The four band configurations, described in Sect. 2, are listed in Cols. 6–11. The references for each cluster can be found in the last column. *Different spectroscopic data sets are described in the text. The case mixed* is similar to the mixed one, with the only difference that it does not include the two clusters A2744 and A370. The cluster RX J2248.7–4431 is also known as Abell S1063.

References. (1) Rosati & Clash-VLT Team (in prep.); (2) Monna et al. (2015); (3) Caminha et al. (2019); (4) Mahler et al. (2018); (5) Caminha et al. (2016); (6) Lagattuta et al. (2019); (7) Grillo et al. (2015), (8) Balestra et al. (2016); (9) Caminha et al. (2017b); (10) Caminha et al. (2017a); (11) Grillo et al. (2016); (12) Treu et al. (2016).
4.2. EXP2: Selection of clusters as blind test set

A second set of experiments was devoted to the study of the CNN capability to predict cluster membership of sources belonging to clusters that are not included in the training set, that is, avoiding having member galaxies belonging to the same cluster populating both the training and test sets. Thus, we considered A370 ($z = 0.375$), MS 2137-2353 (MS2137, $z = 0.316$), and MACS J0329-0211 (M0329, $z = 0.450$) as blind test clusters, while the remaining clusters were organised into three different training sets based on different redshift ranges, as shown in Fig. 3. Specifically:

- Narrow: clusters with redshift $0.332 \leq z \leq 0.412$ (514 CLMs, 555 NCLMs)
- Intermediate: clusters with redshift $0.286 \leq z \leq 0.467$ (898 CLMs, 1157 NCLMs)
- Large: clusters with redshift $0.174 \leq z \leq 0.606$ (1286 CLMs, 1725 NCLMs)

The training set configurations were mostly organised to identify CLMs in A370. This is the most significant test bench as ROC curves, while in Fig. 7, the performances are summarised by combining the results from the four clusters based on their ROC curves (top), the trade-off between purity and completeness with a cross-over at $\sim 88\%$ for blue objects. The results reflect the underlying similarity between blue members and background objects, which implies that they cannot be separated easily. This was confirmed by the analysis of false positives and false negatives discussed in Sect. 6.

4.3. EXP3: Comparison with photometric approaches

This section is dedicated to a comparison of the classification performance of cluster members using the image-based DL method described above along with two different techniques based on photometric catalogues. The first is a random forest classifier (developed by our team) and the second one is a photometry-based Bayesian model described in Grillo et al. (2015) and in Appendix C.

In this experiment, our CNN was trained with the mixed* filter set (see Sect. 2). We focused on the results obtained by these three methods on R2248, M0416, M1206, and M1149. The statistical estimators are shown in detail in Table D.5 and in Fig. D.3 as ROC curves, while in Fig. 7, the performances are summarised by combining the results from the four clusters based on their ROC curves (top), the trade-off between purity and completeness (middle), and the usual statistical estimators (bottom). The photometric techniques show an average efficiency around 86–89%, with some values $\geq 96\%$ for the Bayesian approach, although the F1 scores always remain between 83% and 88%. The CNN confirmed its ability to detect CLMs with an F1 score between 87% and 91%. The upper panel in Fig. 7 shows that globally CNN reaches an AUC of ~94%, which is ~8% higher than the Bayesian method, while exhibiting the sharpest rise and the highest plateau. This means that for the CNN method there is a larger probability range in which the performances remain stable, while for the other methods a fine-tuning of the probability value is needed to balance purity and completeness. Furthermore, CNN reached the best trade-off between purity and completeness with a cross-over at $\sim 89\%$. A summary of the results is shown in the bottom panel of Fig. 7, where the
differences among the CNN and the two photometric methods are measured using the four statistical estimators. The CNN performances were overall near 90% and remained consistently higher than those of photometric-based methods. Finally, we analysed the common predictions among the three methods, both in terms of correctly classified and misclassified sources, separately for CLMs and NCLMs. Such results are graphically represented in Fig. D.2. All three methods share ~76% of their commonalities (i.e. summing of correct and incorrect predictions), of which, ~97% (i.e. 74.6% with respect to the whole set of common sources) were correctly classified. Common true positives and true negatives (i.e. CLMs and NCLMs that have been correctly classified) were ~75%. The CNN and Bayesian method shared the largest fraction of predictions ~90% (of which ~93% were correct) with respect to the joint classification of CNN and RF (~82%); this implied that RF had a significant fraction of uncommon predictions (~14%).

Concerning the misclassified objects, the methods shared ~2% of incorrect predictions, of which: ~1% of CLMs were common false negatives (FNs, i.e. CLMs sources wrongly predicted as NCLMs), while 2.5% were common false positives (FPs, i.e. NCLMs sources wrongly predicted as CLMs). The CNN exhibited the least fraction of misclassifications (about 10%). The CNN showed a percentage of FNs larger than Bayesian (10% versus 7%), which, in turn, had a wider FP rate (11% versus 17%). Therefore, although CNN and Bayesian methods shared a significant fraction of incorrect predictions (85% of common misclassifications, suggesting the existence of a fraction of sources for which the membership is particularly complex for both of them), these two models exhibit a different behaviour: the CNN tended to produce more pure than complete CLMs samples, whereas the Bayesian method showed the opposite, which is in agreement with what is reported in Table D.5.

5. Photometric selection of CLMs
The experiments described in the previous sections are mostly focused on the classification efficiency and limits of the image-based CNN approach and evaluating its dependence from observational parameters such as redshift, number of CLM, photometric band compositions, magnitude, and colour. In this section, we are mainly interested in evaluating the degree of generalisation capability of the trained CNN in classifying new sources as cluster members, a step process that is commonly referred to as run in the ML context.

In particular, we applied the CNN model to the photometrically selected CLMs in R2248, M0416, M1206, and M1149.
The training set was constructed by combining all clusters with the mixed band configuration, using the k-fold approach (see Sect. 3).

Similarly to what was done to build the knowledge base (see Sect. 2), for the run set we used squared cut-outs ~4″ across, centered on the source positions as extracted by SExtractor (Bertin & Arnouts 1996). Thus, the run set was composed by 5269 unknown sources, of which 1286, 1029, 1246, and 1708 were in the FoV of R2248, M0416, M1206, and M1149, respectively.

The CNN identified a total of 372 members with $F814 \leq 25$ mag, which is approximately the magnitude limit of the spectroscopic members (only ~3% of spectroscopic members has $F814 > 25$), with ~46% of candidate CLMs having membership probabilities larger than 90%. The spatial distribution of both spectroscopic and predicted CLMs are shown in Fig. 8, while the magnitude ($F814$) distribution and the colour-magnitude relations ($F606 − F814$ versus $F814$) for both spectroscopic and predicted members are shown in Fig. 9. The magnitude distributions indicate that the CNN was able to complete the spectroscopic CLMs sample down to $F814 = 25$. This was also confirmed by the analysis of the colour-magnitude diagrams, which show that the photometrically identified CLMs complete the spectroscopic red-sequence at $F814 < 25$, emphasising the CNN capability to disentangle CLMs from background objects. We counted also the number of recognised CLMs within, respectively, 1, 2, and 3σ from the median of differential colour ($F606 − F814$)$_{\text{diff}}$.

6. Discussion

One particular aspect that is often addressed when using ML methods is the impact on the classification performances carried by the amount of data available, both in terms of the number of features (photometric bands) and amount of training objects. The EXP1 (see Sect. 4.1) enabled an analysis of the trade-off between the information carried by the imaging bands and the number of samples in the dataset. As reported in Table D.1 and Table D.2.

### Table 2. Statistical performances of the CNN model in EXP2.

|       | Bright | Faint | Redder | Bluer |
|-------|--------|-------|--------|-------|
| pur   | $85.9 \pm 0.4$ | $82.2 \pm 0.8$ | $91.0 \pm 0.5$ | $79.4 \pm 0.9$ |
| compl | $95.2 \pm 0.7$ | $81.4 \pm 1.0$ | $95.2 \pm 0.6$ | $75.7 \pm 1.0$ |
| F1    | $90.3 \pm 0.4$ | $81.7 \pm 0.8$ | $93.1 \pm 0.7$ | $77.6 \pm 0.8$ |

|       | Bright | Faint | Redder | Bluer |
|-------|--------|-------|--------|-------|
| pur   | $88.4 \pm 0.7$ | $83.6 \pm 0.9$ | $90.5 \pm 0.7$ | $79.8 \pm 1.0$ |
| compl | $96.8 \pm 0.7$ | $80.8 \pm 1.2$ | $93.9 \pm 0.4$ | $77.4 \pm 1.2$ |
| F1    | $92.4 \pm 0.7$ | $82.1 \pm 0.9$ | $92.2 \pm 0.8$ | $78.6 \pm 0.9$ |

|       | Bright | Faint | Redder | Bluer |
|-------|--------|-------|--------|-------|
| pur   | $80.7 \pm 0.6$ | $81.1 \pm 1.7$ | $88.3 \pm 0.9$ | $74.4 \pm 1.2$ |
| compl | $98.0 \pm 1.0$ | $85.1 \pm 0.6$ | $95.1 \pm 0.6$ | $78.6 \pm 0.8$ |
| F1    | $89.3 \pm 0.5$ | $83.0 \pm 1.7$ | $91.7 \pm 0.7$ | $76.5 \pm 1.0$ |

|       | Bright | Faint | Redder | Bluer |
|-------|--------|-------|--------|-------|
| pur   | $90.8 \pm 1.0$ | $76.7 \pm 1.5$ | $87.5 \pm 0.3$ | $72.0 \pm 1.3$ |
| compl | $88.9 \pm 1.2$ | $80.0 \pm 0.9$ | $90.6 \pm 0.6$ | $76.2 \pm 1.0$ |
| F1    | $89.7 \pm 1.1$ | $78.3 \pm 1.2$ | $89.0 \pm 0.4$ | $74.1 \pm 1.1$ |

Notes. Best results are emphasised in bold.
there was a small improvement of efficiency (~3%) by increasing the size of the training sample by 34%, when comparing the two mixed and mixed* configurations. However, these two data samples included both optical and infrared information. To better understand this important aspect, we performed a comparison between data samples with and without the infrared bands. Such analysis was carried out by directly comparing the two ACS and ALL configurations, although the sample size of the second one was ~30% smaller. The results, shown in Table D.1 and Table D.2, suggested that the addition of infrared imaging adequately compensated the smaller size of the training set.

We also investigated the dependence of member classification performance on the magnitudes and colours. Here, the EXP2a showed very good performances of the method for both bright and faint sources, although with a slightly lower efficiency in identifying fainter objects. On the other hand, EXP2b showed a mean efficiency of ~91% in classifying red galaxies, which was reduced to ~77% for blue objects (see Table 2).
Fig. 9. CNN membership prediction (run) together with spectroscopic sources, represented as (i) CLMs distribution of $F_{814}$ magnitudes (first row), (ii) differential colour – magnitude sequence for both CLMs and NCLMs. Spectroscopic CLMs are shown in green, candidate members in purple, spectroscopic NCLMs with blue squares and candidate NCLMs with open cyan circle. We only plot CNN cluster members with $F_{814} \leq 25$ mag. The grey region within the CM diagrams limits the area corresponding to $\pm 1\sigma$ from the median (dashed horizontal line) of $(F_{606} - F_{814})_{\text{diff}}$.

Fig. 10. Ensemble of object cut-outs with a size of 64 pixels (~4’’), corresponding to some specific CNN predictions in the clusters R2248 (first three rows) and M0416 (last two rows). The True Positives (TPs), i.e. the CLMs correctly identified, are shown on first and fourth row with green boxes, while False Positive and False Negative (FPs and FNs) are shown on the second, third and fifth row, framed by red and blue boxes, respectively. The images were obtained by combining five HST bands: $F_{435}$, $F_{606}$, $F_{814}$, $F_{105}$, $F_{140}$. The figure shows sources in the $F_{814}$ band with a magnitude $F_{814} \leq 25$ mag. TPs are shown together with their spectroscopic redshift, while FNs together with their cluster rest-frame velocity separation. For convenience, in the case of FPs, their cluster velocity separations are quoted when within $\pm 9000$ km s$^{-1}$, otherwise their redshift is shown.
To further investigate the robustness in the identification of cluster members (i.e. the positive class), from the classification confusion matrices, we defined true positives (TPs) the CLMs correctly classified, false positives (FPs) the NCLMs classified as CLMs, false negatives (FNs) the CLMs classified as NCLMs, and, finally, true negatives (TNs) as NCLMs correctly classified. A short sample of TPs, FPs and FNs in R2248 and M0416, and in M1206 and M1149 are shown in Figs. 10 and 11, respectively.

We explored the model predictions, by inspecting the TPs and the distribution of FPs and FNs as function of their magnitude and colour.

A critical aspect of the classification of members within the central cluster region is the impact of crowding. Therefore, we specifically focused on the DL ability to predict cluster membership in such circumstances (see a few examples of cut-outs in Figs. 10 and 11).

We introduced a contamination index (CI) for each cut-out, defined as: $CI = \frac{1}{N_c} \sum_{i=1}^{N_c} d_i \cdot F_{814}$, where $N_c$ is the number of contaminants in the cut-outs, $d_i$ is the distance in arcsec between the central source and $i$-th contaminant, while $F_{814}$ is the magnitude of the contaminating source. The indices for cut-outs without contaminants were set to zero. Then, we normalised this index in the $[0, 1]$ interval. Figure 12 shows that the four contamination index distributions of, respectively, TPs, TNs, FPs and FNs mostly overlapped and followed the same trend. In fact, the 48% of FNs and 28% of FPs had a non-zero contamination index, as well as the 31% and 43% of TNs and TPs. The lack of a correlation between the contamination index and incorrect prediction rates (FPs and FNs) suggests that the source crowding did not significantly affect the CNN classification efficiency.

By analysing the FP and FN rows in Figs. 10 and 11, we can see an interesting dichotomy: FPs appear as red galaxies, while FNs as blue; in addition, the FPs have $F_{814} < 24$, whereas FNs are found also down to $F_{814} \sim 25$. In order to quantify such behaviours, we analysed the distribution of FPs and FNs in terms of: (i) the $F_{814}$ magnitude for both FPs and FNs (Fig. 13a); (ii) the correlation between the CNN incorrect predictions and differential colours ($F_{606} - F_{814}$) (Fig. 13b). These results are summarised in Table 3.
was not due to any failure on the part of the model, but, rather, to the poor sampling of such objects within the parameter space available to train the model. This was also confirmed when comparing the percentage of FPs and FNs with respect to the percentage of CLMs and NCLMs in Table 3 as a function of magnitude. In fact, Table 3 showed that the model tried to reproduce the distribution in terms of fractions of CLMs for FPs, and in terms of the fraction of NCLMs for FNs.

Finally, we analysed the correlations between the CNN incorrect predictions and colours. These distributions are shown in Fig. 13b using the differential colour \((F606−F814)\text{diff} < −0.5\) while, in Table 3, the misclassification percentages are summarised. Also in this case, the distributions of FPs and FNs as a function of colours, are mimicking, respectively, the distributions of CLMs and NCLMs in Table 3.

Very blue sources \((F606−F814)\text{diff} < −0.5\) populated only 5.8% of CLMs and represented the ~35.4% of incorrect predictions, which is very similar to the fraction of very blue sources in the population of NCLMs (i.e. 43.2%). Conversely, redder sources were typically correctly classified, showing a FN rate of 16.6%. Moreover, from the fraction of FN/CLMs, we observed that almost all the blue cluster members were wrongly classified as NCLMs (see Col. 4 in Table 3 and Fig. 13b).

Regarding FPs, there was not a real classification problem with faint and very blue objects, whose rates in terms of CLMs were, respectively, 3.4% and 5.8%, corresponding to 2.2% and 4.3% of incorrect predictions, respectively. From Table 3, it was also evident that within red misclassifications, FPs were more frequent than FNs (29.5% versus 16.6%), reproducing the distributions of CLMs (39.2%) and NCLMs (15.4%), respectively.

Figure 14 shows the colour-magnitude relation of CLMs (green squares), overlapping the FP (red cross), FN (blue cross) and NCLM (grey circle) distributions. It emphasises the CLMs undersampling of the blue and faint region, together with the large concentration of FNs among bluer and fainter sources (see blue crosses). Among all the FNs, ~35% are very blue \(((F606−F814)\text{diff} < −0.5)\), ~40% of these had \(F814 > 25\) mag, suggesting that in the bluer region the FNs follows the NCLM distribution, while among FPs, ~64% of them are red \(((F606−F814)\text{diff} > −0.1),\) but only ~1% of these have magnitude fainter than \(F814 > 25\) mag. On the other hand, ~35% of all FPs were within the yellow contours, which refer to the \(1\sigma\) colour-magnitude relation, indicating that they were on the red sequence.

In order to understand the impact of this misclassification of faint and very blue sources, we report, in Table 4, the statistical estimators for the stacked sample and, individually for R2248, M0416, M1206, M1149, considering either the whole sample or by removing sources with \(F814 > 25\) and very blue objects, that is, with \((F606−F814)\text{diff} < −0.5\). By comparing these results, we observed a relevant increase of the completeness (for the stacked sample, it goes from 84.8% to 90.8%). This was mainly motivated by the sensible reduction of the FNs amount, which, by definition, had a higher impact on the completeness, rather than on other estimators. In fact, the purity and F1 score showed a smaller improvement, going, respectively, from 87.9% to 88.4% and from 86.3% to 88.9%.

In summary, the FNs were mainly blue and faint. This was expected, given their under-representation in the dataset and their similarity with NCLMs. We note, in fact, that we were mapping a population of cluster members in the central and highest density region of clusters, dominated by a high fraction of bright and red members. Nevertheless, the simple exclusion of fainter sources with \(F814 > 25\) and \((F606−F814)\text{diff} < −0.5\) improved the CNN performance. Similar performances in terms of the distribution of false positives and negatives for sources with \(F814 > 25\) and \((F606−F814)\text{diff} < −0.5\) were obtained by the random forest classifier and the photometry-based Bayesian method. By comparing the behaviour of these three models on four clusters (R2248, M0416, M1206 and M1149), we found that the rate of blue FN is 28% for the Bayesian method and 25% for the random forest versus the 20% for the CNN. The rate of faint FN is 1% for the random forest and 6% for the Bayesian method versus the 5% of CNN. For what concerns FPs, the CNN, being the purest method, preserved the lowest contamination for both bluer and fainter members, with only four NCLMs classified as CLMs, compared with the 12 and 24 NCLMs for the Bayesian method and the random forest, respectively.

This comparison, while it confirms the good performances of the CNN, also shows that the three methods have comparable efficiencies in the faint and blue region of the parameter space, which is likely due to undersampling of members in this region of the knowledge base, as pointed out above. This is due to the fact that the population of galaxies in the densest central cluster regions is brighter and redder than that of the less dense and outer cluster regions (see Annunziatella et al. 2014; Mercurio et al. 2016 for the specific study of M1206). Clearly, an improvement of the model’s performances would require including member galaxies in the outer cluster regions and balancing the number of bluer and fainter members. In our case, even if the spectroscopic data cover more than two cluster virial radii, multi-band HST imaging with sufficient depth is only available in the central cluster regions.

Finally, we used both spectroscopic members and candidate CLMs identified by CNN to estimate the cumulative projected number of cluster members and the differential number density profiles (Fig. 15). According to our previous analysis, we excluded candidate CLMs with \(F814 > 25\) mag, where only ~3% of spectroscopic members were present. To properly compare profiles of clusters with different virial masses, we computed the values \(R_{200}\) from of the values of \(M_{200}\) obtained by Umetsu et al. (2018) with independent weak lensing
measurements\(^3\). We then computed all profiles as a function of the projected radius in units of \(R_{200}\) and rescaled them by the number of members, \(N_0\), found within the radius \(R/R_{200} = 0.15\) in each cluster. In Fig. 15, we showed the cumulative projected number and the differential projected number density profiles of cluster members after applying such renormalisations, where the shaded areas correspond to 68% confidence levels. Interestingly, we found that the radial distributions of all clusters followed a universal profile, including M0416, which is an asymmetric merging cluster. We noted that a similar homology relation among rescaled projected mass profiles was found in Bonamigo et al. (2018) and Caminha et al. (2019), using strong lensing modelling. This result confirms that our methodology was able to identify the CLM population with a high degree of purity and completeness.

7. Conclusions

In this work, we carry out a detailed analysis of CNN capabilities to identify members in galaxy clusters, disentangling them from foreground and background objects, based on imaging data alone. Such a methodology, therefore, avoided the time consuming and challenging task of building photometric catalogues in cluster cores. We used OPT-NIR high quality HST images, supported by MUSE and CLASH-VLT spectroscopic observations of fifteen clusters, spanning the redshift range \(z_{\text{cluster}} = (0.19, 0.60)\).

---

\(^3\) We note again that \(R_{200} = \left( \frac{300}{M_{200}} \right)^{1/3} H(z_{\text{cl}})\), where \(H(z_{\text{cl}})\) is the Hubble constant computed at the cluster redshift.
We used this extensive spectroscopic coverage to build a training set by combining CLMs and NCLMs. We performed three experiments by consecutively varying the HST band combinations and the set of training clusters to study the dependence of DL efficiency on (i) the cluster redshift (EXP1); and (ii) the magnitude and colour of cluster galaxies (EXP2). We also compared the CNN performance with other methods (random forest and Bayesian model), based instead on photometric measurements (EXP3). The main results can be summarised as follows:

- Despite members belonging to clusters spanning a wide range of redshift, the CNN achieved a purity-completeness rate $\geq 90\%$, showing a stable behaviour and a remarkable generalisation capability over a relatively wide cluster redshift range (Sect. 4.1).

- The CNN efficiency was maximised when a large set of sources was combined with HST passbands, including both optical and infrared information. The robustness of the trained model appeared reliable even when a subset of clusters was moved from the training to the blind test set, causing a small drop ($<5\%$) in performance. We observed some performance differences for bright and faint sources, as well as for red and blue galaxies. However, the results maintained the purity, completeness and F1 score greater than $72\%$ (Table 2 in Sect. 4.2).

- By using images, rather than photometric measurements, the CNN technique was able to identify CLMs with the lowest rate of contamination and the best trade-off between purity and completeness, when compared to photometry-based methods, which instead require a critical fine-tuning of the classification probability.

- The false negatives, that is, the NCLMs wrongly classified as CLMs were mainly blue and faint. This was simply the result of their limited under-sampling in the training dataset, as well as their similarity with NCLMs. However, by excluding sources with $F814 > 25$ mag and $(F606 - F814)_{\text{diff}} < -0.5$, the CNN performance improved significantly. These performances reflected the capability of the CNN to classify unknown objects, from which a highly complete and pure magnitude limited sample of candidate CLMs could be extracted for several different applications in the study of the galaxy populations and mass distribution of galaxy clusters via lensing techniques.

Therefore, based on an adequate spectroscopic survey of a limited sample of clusters as a training base, the proposed methodology can be considered a valid alternative to photometry-based methods, circumventing the time-consuming process of multi-band photometry, and working directly on multi-band imaging data in counts. To improve CNN performance to recognise the faintest and blue CLMs, it would be desirable to plan both HST and spectroscopic observations also covering control fields in the outer cluster regions, with the same depth and passbands as the central regions.

Furthermore, the generalisation capability of this kind of models makes them both versatile and reusable tools. In fact, the convolution layers of a trained deep network can be reused as shared layers in larger models, such as the Faster Region CNN (Ren et al. 2015) and Masked Region CNN (He et al. 2017), which exploit kernel weights to extract multidimensional information suitable to performing object detection. Such architectures have already found interesting astrophysical applications, for example, in the identification of radio sources (Wu et al. 2019) and the automatic deblending of astronomical sources (Burke et al. 2019).

In future works, we will extend this analysis to wide-field ground-based observations and explore other promising deep learning architectures, such as deep auto-encoders (Goodfellow 2010) and conditional generative adversarial networks (Mirza & Osindero 2014), to integrate the ground-based lower resolution images with the high quality of HST images.

Fig. 14. Colour-magnitude relation for the CLMs (green squares), with the overlapped distributions of FPs (red crosses), FNs (blue crosses) and NCLMs (grey circles), for the sample of fifteen clusters (stacked). The yellow contour delimits the red-sequence at $1\sigma$ confidence level. Colours reported on the y-axis are corrected for the mean red-sequence of each cluster (see Sect. 4).
in cluster fields. We also plan to investigate new techniques to overcome the problem of missing data, thus increasing the size of the training set with a more homogeneous sampling of the entire parameter space.

Acknowledgements. The authors thank the anonymous referee for the very useful comments and suggestions. The software package HIGHCOOLs (Hierarchical Generative Hidden Convolution Optimization System, http://dame.oacn.inaf.it/highcools.html), developed within the DAME project (Brescia et al. 2014), has been used for the deep learning models described in this work. We acknowledge funding by PRIN-MIUR 2017/2W3CC/32 “Zooming into dark matter and proto-galaxies with massive lensing clusters”, INAF main stream 1.05.01.86.20: “Deep and wide view of galaxy clusters (ref. Mario Nonino)” and INAF mainstream 1.05.01.86.31 (ref. Eros Vanzella). MB acknowledges financial contributions from the agreement ASI/INAF 2018-23-H.0, Euclid ESA mission – Phase D and with AM the INAF PRIN-SKA 2017 program 1.05.01.88.04. CG acknowledges support through Grant no. 10123 of the VILLUM FONDEN Young Investigator Programme. In this work several public software was used: Topcat (Taylor 2005), Astropy (Astropy Collaboration 2016, Astropy Collaboration (Robitaille, T. P., et al.) 2013, A&A, 558, A33 Astropy Collaboration (Price-Whelan, A. M., et al.) 2018, AJ, 156, 123 Bacon, R., Vernet, J., Bosisio, E., et al. 2014, Messenger, 157, 13 Balestra, I., Mercurio, A., Sartoris, B., et al. 2016, ApJS, 224, 33 Batista, G. E. A. P. A., & Monard, M. C. 2003, Appl. Artif. Intell., 17, 519 Bengio, Y. 2012, Neural networks: Tricks of trade, Springer, 437 Bergamini, P., Rosati, P., Mercurio, A., et al. 2019, A&A, 631, A130 Bertin, E., & Arnouts, S. 1996, Ap&SS, 117, 393

References

Abadi, M., Agarwal, A., Barham, P., et al. 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, software available from tensorflow.org (San Francisco: Astronomical Society of the Pacific).

Annunziatella, M., Biviano, A., Mercurio, A., et al. 2014, A&A, 571, A80 Annunziatella, M., Mercurio, A., Biviano, A., et al. 2016, A&A, 585, A160 Annunziatella, M., Bonamigo, M., Grillo, C., et al. 2017, ApJ, 851, 81 Astropy Collaboration (Robitaille, T. P., et al.) 2013, A&A, 558, A33 Astropy Collaboration (Price-Whelan, A. M., et al.) 2018, AJ, 156, 123 Bacon, R., Vernet, J., Bosisio, E., et al. 2014, Messenger, 157, 13 Balestra, I., Mercurio, A., Sartoris, B., et al. 2016, ApJS, 224, 33 Batista, G. E. A. P. A., & Monard, M. C. 2003, Appl. Artif. Intell., 17, 519 Bengio, Y. 2012, Neural networks: Tricks of trade, Springer, 437 Bergamini, P., Rosati, P., Mercurio, A., et al. 2019, A&A, 631, A130 Bertin, E., & Arnouts, S. 1996, Ap&SS, 117, 393

Bishop, C. M. 2006, Pattern Recognition and Machine Learning (Information Science and Statistics) (Secaucus, NJ, USA: Springer-Verlag, New York, Inc.)

Biviano, A., Rosati, P., Balestra, I., et al. 2013, A&A, 558, A1 Bonamigo, M., Grillo, C., Ettori, S., et al. 2018, ApJ, 864, 98 Breiman, L. 2001, Mach. Learn., 45, 5 Brescia, M., Cava, G., D’Abrusco, R., Longo, G., & Mercurio, A. 2013, ApJ, 772, 140 Brescia, M., Cava, G., Longo, G., et al. 2014, PASP, 126, 783 Brescia, M., Cava, G., Amaro, V., et al. 2018, in Data Analytics and Management in Data Intensive Domains, eds. L. Kalimichchenko, Y. Manolopoulos, O. Malkov, et al. (Cham: Springer International Publishing), Commun. Comput. Inf. Sci., 822, 61 Burke, C. J., Aleo, P. D., Chen, Y.-C., et al. 2019, MNRAS, 490, 3952 Caminha, G. B., Grillo, C., Rosati, P., et al. 2016, A&A, 587, A80 Caminha, G. B., Grillo, C., Rosati, P., et al. 2017a, A&A, 600, A80 Caminha, G. B., Grillo, C., Rosati, P., et al. 2017b, A&A, 607, A93 Caminha, G. B., Rosati, P., Grillo, C., et al. 2019, A&A, 632, A36 Cappellari, M., Scott, N., Atlante, K., et al. 2013, MNRAS, 432, 1709 Cavuoti, S., Brescia, M., De Stefano, V., & Longo, G. 2015, Exp. Astron., 39, 45 Chollet, F., et al. 2015, Keras, https://keras.io Coe, D., Umetu, K., Zitrin, A., et al. 2012, ApJ, 757, 22 Cui, X., Goel, V., & Kingsbury, B. 2015, IEEE/ACM Trans. Audio Speech Lang. Process., 23, 1469 Devroye, L., Györfi, L., & Lugosi, G. 1996, in A Probabilistic Theory of Pattern Recognition, (Springer), Stochastic Modell. Appl. Probab., 31, 1 Dieamond, J., & Moore, B. 2011, Adv. Sci. Lett., 4, 297 D’Isanto, A., Cavuoti, S., Brescia, M., et al. 2016, MNRAS, 457, 3119 Duch, J., Hazan, E., & Singer, Y. 2011, J. Mach. Learn. Res., 12, 2121 Girardi, M., Mercurio, A., Balestra, I., et al. 2015, A&A, 579, A4 Goodfellow, I. J. 2010, Technical Report: Multidimensional, Downsampling Convolution for Autoencoders, Tech. rep. (Université de Montréal)

Goodfellow, I., Bengio, Y., & Courville, A. 2016, Deep Learning (MIT Press), http://www.deeplearningbook.org Grillo, C., Suyu, S. H., Rosati, P., et al. 2015, ApJ, 800, 38 Grillo, C., Karman, W., Suyu, S. H., et al. 2016, ApJ, 822, 78 Hanley, J. A., & McNeil, B. J. 1982, Radiology, 143, 29 Hastie, T., Tibshirani, R., & Friedman, J. 2009, in The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, (New York: Springer), Springer Ser. Stat. He, K., Zhang, X., Ren, S., & Sun, J. 2015, ArXiv e-prints [arXiv:1512.03385]
Appendix A: Methodology

Fig. A.1. Data augmentation example for a CLM at redshift $z = 0.531$ (e.g. within the gravitational potential of M1149). Five HST bands are represented from the top to the bottom ($F435$, $F606$, $F814$, $F105$, $F140$). The first column shows the original cut-out, while the three rotations (90°, 180°, 270°) are reported inCols. 2–4. The two vertical and horizontal flips are shown in the last two columns.

The data preparation phase, preceding the application of the ML based classifiers, is organised as a series of pre-processing steps, detailed in the following sections.

A.1. Data augmentation

The cut-outs have been rotated around the three right angles and flipped with respect to the horizontal and vertical axes (an example of such process is shown in Fig. A.1). Given the considerable number of model parameters to fit ($\sim 10^5$), deep learning networks require an adequate amount of samples, in order to avoid overfitting (Cui et al. 2015; Perez & Wang 2017). However, an uncontrolled augmentation could introduce false correlations among the training samples. Therefore, only a fraction of sources have been subject to these transformations: 15% of the available images have been randomly extracted and used for such transformations mentioned above. The resulting augmentation factor was 1.75 times the original dimension of the training set. Obviously, such augmentation process involved only the training images.

A.2. Setup of training and test sets

Before the k-fold splitting and the augmentation process, described above, we randomly extracted a small sample of sources (10% of the data set), reserved as validation set during the training phase in order to control the gradual reduction of the learning rate on the plateau of the cost function (Bengio 2012) and an early stopping regularisation process (Prechelt 1997; Raskutti et al. 2011). The data preparation flow is depicted in Fig. A.2: (i) the dataset is composed by multi-bands images; (ii) a fraction of sources (10%) is extracted as validation set; (iii) the remaining samples are split into $k = 10$ folds without overlapping; (iv) for each of them, a fraction (15%) of samples is augmented through cut-out rotations and flips; (v) the training sets are built by concatenating $k - 1$ folds (composed by the original images and the artefacts) and the learning is evaluated on the $k$th fold (without artefacts), acting as blind test; (vi) finally, the model performances are evaluated on the whole training set, obtained by stacking all its (test) folds.

A.3. Statistical evaluation of performance

In order to assess the model classification performances, we chose the following statistical estimators: “average efficiency” (among all classes, abbreviated as “AE”), “purity” (also known as “positive predictive value” or “precision”, abbreviated as “pur”), “completeness-“ (also known as “true positive rate” or “recall”, abbreviated as “comp”), and F1-score (a measure of the combination of purity and completeness, abbreviated as “F1”).

In a binary confusion matrix, as in the example shown in Table A.1, columns indicate the class objects as predicted by the classifier, while rows refer to the true objects per class. The main diagonal terms contain the number of correctly classified objects for each class, while the terms FP and FN report the amount of, respectively, false positives and false negatives. Therefore, the derived estimators are computed as:

\[
\begin{align*}
\text{AE} &= \frac{TP + TN}{TP + FP + TN + FN} \\
\text{pur} &= \frac{TP}{TP + FP} \\
\text{comp} &= \frac{TP}{TP + FN} \\
F1 &= 2 \cdot \frac{\text{pur} \cdot \text{comp}}{\text{pur} + \text{comp}}
\end{align*}
\]

The AE is the ratio between the sum of the correctly classified objects (for all the involved classes) and the total amount of objects; it describes an average evaluation weighted on all involved classes. The “purity” of a class is the ratio between the correctly classified objects and the sum of all objects assigned to that class (i.e. the predicted membership); it measures the precision of the classification. The “completeness” of a class is the ratio between the correctly classified objects and the total amount of objects belonging to that class (i.e. the “true” membership), it estimates the sensitivity of the classification. Finally, the F1-score is the harmonic average between purity and completeness. By definition, the dual quantity of purity is the “contamination”, a measure which indicates the amount of misclassified objects for each class.

Moreover, from the probability vector (i.e. the set of values stating the probability that an input belongs to a certain class), it is possible to extract another useful estimator, the receiver

| Predictions | Positive | Negative |
|-------------|----------|----------|
| True Positive | $TP$ | $FN$ |
| False Positive | $FP$ | $TN$ |

Notes. In a confusion matrix, columns indicate the number of objects per class, as predicted by the classifier, while rows are referred to the true (known) objects per class. Hence, the main diagonal terms report the number of correctly classified objects for each class. While, the terms FP and FN count, respectively, the false positives and false negative quantities.

Table A.1. Generic confusion matrix for a binary classification problem.
The operating characteristic (ROC) curve is a diagram in which the true positive rate is plotted versus the false positive rate by varying a membership probability threshold (see Fig. 6). The overall score is measured by the area under the ROC curve (AUC), where an area of 1 represents a perfect classification, while an area of 0.5 indicates a useless result (akin to a toss of a coin).

Appendix B: Convolution neural networks

In this appendix, CNNs theory and our specific implementation are briefly described, while a synthetic view of the implemented model is shown in Fig. B.1.

As any other artificial neural networks, convolution neural networks (CNNs, LeCun et al. 1989) are inspired by biological models. Artificial neurons are arranged in several layers, where each neuron takes as input the signal coming from neurons belonging to the previous layer; such as biological neurons, the variation of the synaptic connection sensibility (with respect to a certain input signal) is correlated to the learning mechanism (Hebb 1949). During the training, these connection sensibilities among layers (i.e. the weights) are adapted through a forward-backward mechanism, at the base of the iterative learning process (Bishop 2006). After training, supervised Machine Learning methods define a non-linear relation between the input and output spaces, which is encoded within the weight matrices.

CNNs represent one of the most widely-used supervised techniques among the Deep Neural Networks (DNN, Goodfellow et al. 2016), whose peculiarity is an ensemble of receptive fields which trigger neuron activity. The receptive field is represented by a small matrix (called as kernel or filter), which connects two consecutive layers through a convolution operation. Similar to the adaptation mechanism imposed by supervised machine learning, the kernels are modified during the training. The peculiarity of such kind of models is the capability to automatically extract meaningful features from images (such as edges and shapes), which become the input vector to any standard ML model that outputs the class of the input image. The idea behind CNN is a convolution subsampling chain mechanism: deep networks are characterised by tens of layers (in some cases hundreds, as proposed by He et al. 2015 and Xie et al. 2016), where at each depth level, the convolution acts as a filter, emphasising (or suppressing) some properties; while the subsampling (often called pooling) makes sure that only essential information is moved towards the next layer.

CNNs are organised as a hierarchical series of layers, typically based on convolution and pooling operations. Convolution kernel is represented by a 4-D matrix \( K \), where the element \( K_{i,j,k} \) is the connection weight between the output unit \( i \) and the input unit \( j \), with an offset of \( k \) rows and \( l \) columns. This kernel is convoluted with the input signal and adapted during the training. Given an input \( V \), whose element \( V_{i,j,k} \) represents an observed data value of the channel \( i \) at row \( j \) and column \( k \), the neuron activity can be expressed as (Goodfellow 2010):

\[
Z_{i,j,k} = c(K, V, s)_{i,j,k} + b
\]

\[
= \sum_{m,n} V_{(i,j,l)_{m,n}} K_{i,j,m,n} + b \tag{B.1}
\]

\[
Z_{i,j,k} \leftarrow p(Z, d)_{i,j,k} \tag{B.2}
\]

\[
Z_{i,j,k} \leftarrow f(Z, \{a_l\})_{i,j,k} \tag{B.3}
\]

where \( c(K, V, s) \) is the convolution operation between the input \( V \) and the kernel \( K \) with stride \( s \); \( b \) is an addend that acts as bias; \( p(Z, d) \) is the pooling operation with down-sampling factor \( d \); \( f(Z, \{a_l\}) \) is the activation function characterised by the set of hyper-parameters \( \{a_l\} \). The pooling function reduces the dimension, by replacing the network output at a certain location with a summary statistic of nearby outputs (Goodfellow et al. 2016).

Unlike traditional artificial neural networks (e.g. Multi-Layer Perceptron), where all neurons of two consecutive layers are fully connected among them, the connection among neurons in a CNN is “sparse”, that is, the interaction between neurons belonging to different layers is limited to a small fraction. This reduces the number of operations, the memory requirements and, thus, the computing time. The output layer consists of an ensemble of sub-images with reduced dimensions, called feature maps, each of them represents a feature extracted from the original signal, processed by the net in order to solve the assigned problem.
Another common operation performed during the training of a CNN is the random dropout of weights. This function prevents units from co-adapting, reduces significantly overfitting and gives major improvements over other regularization methods (Srivastava et al. 2014). At the end of the network, the resulting feature maps ensemble is fully connected with one or more hidden layers (also called “dense layers”), the last of which, in turn, is fully connected to the output layer. The net output must have the same shape of the known target: within the supervised learning paradigm, the comparison between output and target induces the kernel adaptations. When the net task is a classification problem (as in this work), the output is a matrix of probabilities, that is, each sample has a membership probability related to any class of the problem. In order to disentangle objects in the train set, minimising a loss function (or cost function). The most common choice for the loss function is the cross-entropy (Goodfellow et al. 2016):

$$C(y, \bar{y}) = \sum_{j=1}^{n_{classes}} y_j \ln \bar{y}_j + (1 - y_j) \ln (1 - \bar{y}_j) \tag{B.4}$$

where $y$ is the target and $\bar{y}$ is the output of the final layer. Thus, during the training, images extracted from the train set are propagated through the network, while weights and biases are adapted along with a backward flow in order to minimise the cost function. To perform such optimization, it is necessary to state the minimization algorithm. The simplest and most used optimiser is the Stochastic Gradient Descent (Bishop 2006), but in recent years several optimisers have been proposed (e.g. Duch et al. 2011; Zeiler 2012; Kingma & Ba 2014), which offer a faster convergence to the minimum, avoiding the local dump of the cost function. In this work we chose Adadelta (Zeiler 2012) as optimiser. Furthermore, we included (i) an early stopping regularisation criterion (Prechelt 1997; Raskutti et al. 2011), preventing overfitting; and (ii) a gradual reduction of the learning rate on the plateau of the loss function (as function of epochs, Bengio 2012). Both techniques exploit a validation set, extracted from the train set, used to compute and evaluate the learning efficiency within the training cycles. In our case, to avoid memory loss, the network has been trained with input data batches of size equals to 64 patterns.

The architecture of our VGGNET-like model is reported in Table B.1. It is composed of 47 layers and convolution kernels with a window size of $3 \times 3$. The max pooling criterion was preferred to the average algorithm, in order to reduce the noise contribution. We set the Leaky version of a Rectified Linear Unit (LeReLU, Maas et al. 2013) as activation function for all the neurons. This type of activation allows (i) a small, non-zero gradient also when the unit is saturated and not active, (ii) a gain of the convergence with the increase of the units, defined as:

$$\text{out}(x) = \begin{cases} 
\alpha \cdot x & x < 0 \\
x & x \geq 0
\end{cases} \tag{B.5}$$

where $\alpha$ is a hyper-parameter set to 0.3.

This network has been implemented through keras (Chollet 2015), with tensorflow (Abadi et al. 2015) as backend system. Both of them are open-source Python libraries, allowing the automatic handling of the Graphic Processing Unit (GPU), achieving a huge gain in terms of computational cost (~700 see Simard et al. 2005). In this work the experiments were performed with an NVIDIA GPU Titan Xp and an NVIDIA GPU Quadro P5000, requiring ~30 minutes to complete the training (on a single fold, see Appendix A.2).
Table B.1. VGGNET-like model configuration.

| Layer               | Output shape | Params # |
|---------------------|--------------|----------|
| Input layer         | (64, 64, NC) | 0        |
| Conv2D              | (64, 64, 64) | 6976     |
| Leaky ReLU          | (64, 64, 64) | 0        |
| Conv2D              | (64, 64, 64) | 36928    |
| Leaky ReLU          | (64, 64, 64) | 0        |
| Max Pool2D          | (32, 32, 64) | 0        |
| Conv2D              | (32, 32, 128)| 73856    |
| Leaky ReLU          | (32, 32, 128)| 0        |
| Conv2D              | (32, 32, 128)| 147584   |
| Leaky ReLU          | (32, 32, 128)| 0        |
| Conv2D              | (16, 16, 256)| 295168   |
| Leaky ReLU          | (16, 16, 256)| 0        |
| Conv2D              | (16, 16, 256)| 590080   |
| Leaky ReLU          | (16, 16, 256)| 0        |
| Conv2D              | (16, 16, 256)| 590080   |
| Leaky ReLU          | (16, 16, 256)| 0        |
| Max Pool2D          | (8, 8, 256)  | 0        |
| Conv2D              | (8, 8, 512)  | 1180160  |
| Leaky ReLU          | (8, 8, 512)  | 0        |
| Conv2D              | (8, 8, 512)  | 2359808  |
| Leaky ReLU          | (8, 8, 512)  | 0        |
| Conv2D              | (8, 8, 512)  | 2359808  |
| Leaky ReLU          | (8, 8, 512)  | 0        |
| Conv2D              | (8, 8, 512)  | 2359808  |
| Leaky ReLU          | (8, 8, 512)  | 0        |
| Max Pool2D          | (4, 4, 512)  | 0        |
| Conv2D              | (4, 4, 512)  | 2359808  |
| Leaky ReLU          | (4, 4, 512)  | 0        |
| Conv2D              | (4, 4, 512)  | 2359808  |
| Leaky ReLU          | (4, 4, 512)  | 0        |
| Max Pool2D          | (2, 2, 512)  | 0        |
| Flatten             | (2048)       | 0        |
| Dense               | (4096)       | 8392704  |
| Leaky ReLU          | (4096)       | 0        |
| Dropout             | (4096)       | 0        |
| Dense               | (4096)       | 16781312 |
| Leaky ReLU          | (4096)       | 0        |
| Dropout             | (4096)       | 0        |
| Dense               | (2)          | 8194     |
| Output Layer        | (2)          | 0        |

Notes. The columns specify the layer operation, the shape of the output and the number of parameters to fit. The output shape of a layer is a 4-D matrix, but, since the first dimension is the fixed size of the input data batch (with a size of 64 patterns), we do not mention this number to prevent confusion. The total amount of trainable parameters is larger than 45M. The last dimension of the input layer is the involved number of channels (i.e. the number of photometric bands used), a quantity depending on the specific experiment (see Sect. 2).

Appendix C: Benchmark methods

We compared CNN performances with two techniques based on photometric catalogues: a random forest (RF, Breiman 2001) and a Bayesian Method (briefly described in Grillo et al. 2015).

A Bayesian classifier is a model able to minimise the error probability (Devroye et al. 1996), defined as: \( L(g) = P(g(X) \neq Y) \), where \( (X, Y) \) are pair values in \( \mathbb{R}^d \times \{1, \ldots, M\} \) (i.e. \( Y \) is the ensemble of class labels related to the manifold \( X \)), \( g \) is a classifier (i.e. a function \( g : x \in X \subseteq \mathbb{R}^d \rightarrow y \in \{1, \ldots, M\} \)), \( L \) is an application mapping \( g \) into probabilities. The minimal probability error is denoted \( L^* = L(g^*) \), that can be written as:

\[
g^* = \arg\min_{g: \mathbb{R}^d \rightarrow \{1,...,M\}} P(g(X) \neq Y)
\]

Given a classical linear model \( \bar{y}_i = \sum_{j=1}^p x_j \bar{\theta}_j, i = 1,\ldots,n \), the method estimates \( \{\theta_i\}^n \) in order to minimise a coherent combination of the residuals \( r_i = y_i - \bar{y}_i \). The implemented method exploits a minimum covariance determinant method (Rousseeuw 1984), which is based on the minimization of the median of squared residuals.

Random forest is a machine learning classifier consisting of a collection of tree-structured classifiers \( \{h(x, \theta_k), k = 1,\ldots,K\} \) where the \( \{\theta_k\} \) are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input \( x \). The generalisation error for this algorithm depends on the strength of single trees and from their correlations through the raw margin functions. To improve the model accuracy by keeping trees strength, the correlation between trees is decreased and bagging with a random selection of features is adopted. Bagging, or bootstrap aggregating, is a method designed to improve the stability and accuracy of machine learning algorithms. It also reduces variance and helps to avoid overfitting. In this work, we used the RF provided by Scikit-Learn python library (Pedregosa et al. 2011).

Appendix D: Technical descriptions of our performed experiments

In this appendix, we report tables and figures describing in detail the analysis performed for each experiment.

D.1. EXP1

Table D.1. CNN percentage performances in the EXP1 experiment.

| Class       | %   | mixed | ACS  | ALL | mixed* |
|-------------|-----|-------|------|-----|--------|
| AE          | 86.7| 87.4  | 87.7 | 89.3|
| pur         | 83.1| 85.0  | 86.4 | 88.3|
| CLM compl   | 88.4| 88.5  | 86.4 | 86.7|
| F1          | 85.6| 86.7  | 86.4 | 87.4|
| NCLM compl  | 85.5| 86.7  | 88.9 | 91.2|
| F1          | 87.7| 88.3  | 88.9 | 90.6|

Notes. The performances are related to the four band configurations (see Sect. 2) and expressed in terms of the statistical estimators described in Appendix A.3. The overall best results are highlighted in bold.
With this experiment, we evaluated the CNN capabilities to identify CLMs at different cluster redshifts, \( z_{\text{cluster}} \in (0.2, 0.6) \), using different HST band combinations (see Sects. 2 and 4.1). Furthermore, in this experiment, we studied the dependence on redshift and on the number of spectroscopic sources involved in the training. The results related to this experiment have been summarised in Sect. 4.1.

In Table D.1, we report the results achieved globally by CNN, that is, by combining the available clusters (see also Fig. 4), while Table D.2 outlines the performances for each involved cluster, varying the band combinations. The experiment has been carried out with the \( k \)-fold approach, stacking sources in the FoV of 13 (15 only for the \textit{mixed} configuration) clusters, ensuring that the \( k \)-est fold is populated by objects extracted from each involved cluster, proportionally to the number of available spectroscopic sources, that is, providing adequate coverage of the training set respect to the test set.

The comparison between the band configurations is also shown in Fig. D.1, in which performances and their fluctuations are displayed as function of the involved number of samples. For each configuration, we split the knowledge space into ten disjointed subsets, which have been progressively merged in order to build a dataset with which CNN has been trained and tested, always using the \( k \)-fold approach.

In order to analyse the dependence on redshift, we split the CLM redshift range into five equally populated bins and, to complete the knowledge space, with extracted without repetitions from the NCLM population an appropriate number of objects. The network has been trained within each ensemble adopting the \( k \)-fold approach, using only the \textit{mixed} band combination. The result is graphically shown in Fig. 5 and it is stored in Table D.3, in which, we have specified the fluctuation of estimators as an error estimated on the ten folds.

\section*{D.2. EXP2}

In this experiment, we explored the limits of the CNN in terms of its classification efficiency. With this aim, we excluded three clusters from the training sample, respectively, A370 (\( z = 0.375 \)), MS2137 (\( z = 0.316 \)) and M0329 (\( z = 0.450 \)), which were considered as the blind test set. Such experiment is particularly suitable to evaluate the model capability to predict the cluster membership of sources extracted from clusters unused during training. Furthermore, in this experiment, we varied the training configuration based on three redshift ranges centered on A370 cluster redshift (named as \textit{narrow}, \textit{intermediate} and \textit{large}, see Sect. 4.2 and Fig. 3), exploiting the \textit{mixed} band configuration. This experiment has been described in Sect. 4.2. Table D.4 outlines the results achieved on the three configurations together with Fig. 6. As second step, we analysed the CNN classification capabilities by separating, respectively, brighter from fainter (\textit{EXP2a}), and redder from bluer objects (\textit{EXP2b}). Concerning the magnitude threshold, we split the CLM F814 distribution into two equal-sized sets (F814 limits are 22.0 mag, 21.7 mag, and 21.6 mag for, respectively, A370, M0329 and MS2137). Regarding the colour split, we exploited the correlation between the Balmer break and the differential colour, as shown in Girardi et al. (2015): 
\[
(F814 - F160)_{\text{diff}} = (F814 - F160)_{\text{obs}} - CM(F814),
\]
that is, the difference between the observed colour and the one of the colour-magnitude (CM) relation at a given magnitude. For each cluster, we fitted the CM sequence using a robust linear regression (Cappellari et al. 2013) involving spectroscopic confirmed members. By applying this kind of correction to the source colour, redder members were centered around zero, while bluer objects have differential colours around \( -0.2 \) mag. The differential colour thresholds were \(-0.160\), \(-0.165\), and \(-0.157 \) mag for, respectively, A370, M0329 and MS2137. For both experiments, we opted for a \textit{large} ensemble and \textit{mixed} band configuration. The results are shown in Table 2.

\section*{D.3. EXP3}

This test was devoted to the comparison of CNN performance with two different photometry-based methods, exploiting a random forest classifier (Breiman 2001) and a Bayesian model (Grillo et al. 2015). Both techniques critically use multi-band photometric information, for example, magnitudes and colours. This experiment has been outlined in Sect. 4.3. The Bayesian method has already been applied in order to enlarge the cluster member selection, including galaxies without spectroscopic information, for four clusters: R2248, M0416, M1206, and M1149 (Grillo et al. 2015; Caminha et al. 2016, 2017a; Treu et al. 2016). We compare these methods with our CNN, trained with the \textit{mixed} band configuration, constraining the results to these four involved clusters. The comparison is summarised in Table D.5 in term of statistical estimators, whereas, in Fig. 7 and Fig. D.3, it is shown in terms of ROC curves (see Appendix A.3), in Fig. D.2 in terms of commonalities among predictions. Particularly, we also compared performances between CNN and photometric methods by computing the differences:
\[
\Delta_{\text{estim}} = \text{estim}_{\text{CNN}} - \max\{\text{estim}_{\text{RF}}, \text{estim}_{\text{Bayesian}}\}
\]
for estim \( \in \{ \text{pur, compl, F1, AE} \} \), that is, the difference between CNN metrics and the corresponding maximum scores achieved by RF or Bayesian model. All these differences are listed in the last column of Table D.5, together with the average among these \( \Delta s \) for each cluster (rows \( \mu_s \)).

An additional comparison of the three methods based on common membership predictions (see Fig. D.2), is discussed in Sect. 4.3.
Table D.2. CNN percentage performances evaluated for each cluster and for each band configuration related to the EXP1 experiment.

|        | A383 $z = 0.188$ | R2129 $z = 0.234$ | A2744 $z = 0.308$ |
|--------|------------------|-------------------|------------------|
|        | mixed ACS ALL mixed* | mixed ACS ALL mixed* | mixed ACS ALL mixed* |
| AE     | 77.0 81.8 78.3 83.0 | 89.7 91.6 93.7 92.3 | 93.6 |
| pur    | 77.2 82.9 82.5 86.3 | 76.5 84.6 86.5 84.4 | 95.3 |
| CLM compl | 81.5 82.9 75.0 81.5 | 90.7 91.7 88.9 88.4 | 86.5 |
| F1     | 79.3 82.9 78.6 83.8 | 83.0 88.0 87.7 86.4 | Only mixed* 90.7 |
| pur    | 76.7 80.6 74.4 79.6 | 96.2 95.6 96.2 95.5 | 92.8 |
| NCLM compl | 71.7 80.6 82.1 84.8 | 89.3 91.6 95.3 93.8 | 97.6 |
| F1     | 74.2 80.6 78.0 82.1 | 92.6 93.6 95.8 94.6 | 95.2 |
|        | MS2137 $z = 0.316$ | R2248 $z = 0.346$ | M1931 $z = 0.352$ |
|        | mixed ACS ALL mixed* | mixed ACS ALL mixed* | mixed ACS ALL mixed* |
| AE     | 83.7 81.5 88.2 88.4 | 89.5 86.5 90.2 88.1 | 93.3 |
| pur    | 80.0 79.7 85.7 89.7 | 88.6 85.2 90.7 88.3 | 93.3 |
| CLM compl | 87.8 81.0 90.9 85.4 | 92.4 91.3 92.5 89.8 | 80.6 78.0 75.8 |
| F1     | 83.7 80.3 88.2 87.5 | 90.5 88.1 91.6 89.1 | 77.8 82.9 80.7 86.2 |
| pur    | 87.8 83.1 90.9 87.2 | 88.3 89.5 87.9 87.9 | 80.8 86.4 85.7 85.4 |
| NCLM compl | 80.0 81.9 85.7 91.1 | 86.1 80.7 87.2 86.1 | 95.5 89.9 89.7 100.0 |
| F1     | 83.7 82.5 88.2 89.1 | 88.3 84.3 88.3 87.0 | 87.5 88.1 87.6 92.1 |
|        | M1115 $z = 0.352$ | A370 $z = 0.375$ | M0416 $z = 0.397$ |
|        | mixed ACS ALL mixed* | mixed ACS ALL mixed* | mixed ACS ALL mixed* |
| AE     | 88.1 84.9 89.6 92.5 | 88.9 90.3 90.0 91.5 | 92.2 |
| pur    | 85.7 82.5 90.9 91.8 | 85.8 92.4 90.3 95.7 | 93.3 |
| CLM compl | 93.0 89.5 89.3 94.4 | 87.6 87.1 88.8 86.8 | 87.1 |
| F1     | 89.2 85.8 90.1 93.1 | 86.7 89.7 89.5 91.0 | 91.5 |
| pur    | 91.2 87.9 88.2 93.4 | 89.5 88.6 89.7 88.1 | 89.0 |
| NCLM compl | 82.5 80.0 90.0 90.5 | 87.9 93.3 91.1 96.1 | 96.9 |
| F1     | 86.7 83.8 89.1 91.9 | 88.6 90.9 90.4 91.9 | 92.8 |
|        | M1206 $z = 0.439$ | M0329 $z = 0.450$ | R1347 $z = 0.451$ |
|        | mixed ACS ALL mixed* | mixed ACS ALL mixed* | mixed ACS ALL mixed* |
| AE     | 87.7 90.3 87.4 89.7 | 81.6 81.9 83.3 85.0 | 91.2 90.7 89.7 89.9 |
| pur    | 83.7 89.8 84.2 89.9 | 76.9 76.8 79.1 83.3 | 79.7 81.6 80.4 81.0 |
| CLM compl | 89.7 90.2 86.7 86.5 | 89.6 91.5 88.3 91.0 | 100.0 96.9 93.8 92.2 |
| F1     | 86.6 90.0 85.4 88.2 | 82.8 83.5 83.5 87.1 | 88.7 88.6 86.5 86.2 |
| pur    | 91.3 90.8 89.9 89.6 | 87.9 89.5 88.1 90.0 | 100.0 97.9 96.2 95.6 |
| NCLM compl | 86.2 90.4 88.0 92.3 | 73.9 72.3 78.8 78.3 | 86.6 87.0 87.5 88.7 |
| F1     | 88.7 90.6 88.9 90.9 | 80.3 80.0 83.2 83.7 | 92.8 92.2 91.7 92.0 |
|        | M1311 $z = 0.494$ | M1149 $z = 0.542$ | M2129 $z = 0.587$ |
|        | mixed ACS ALL mixed* | mixed ACS ALL mixed* | mixed ACS ALL mixed* |
| AE     | 77.1 82.5 75.8 78.1 | 85.9 90.7 88.0 89.4 | 85.5 84.6 89.1 91.3 |
| pur    | 72.7 80.3 75.0 76.0 | 74.5 83.3 80.5 82.3 | 85.9 87.3 91.0 93.3 |
| CLM compl | 85.1 77.8 78.3 80.9 | 94.5 92.6 91.5 91.3 | 82.7 83.1 75.3 77.8 |
| F1     | 78.4 79.0 76.6 78.4 | 83.3 87.7 85.6 86.6 | 84.3 85.2 82.4 84.0 |
| pur    | 82.9 84.1 76.7 80.4 | 96.1 95.6 94.0 94.5 | 85.3 85.6 81.0 82.7 |
| NCLM compl | 69.4 86.0 73.3 75.5 | 80.8 89.7 85.7 88.3 | 88.0 89.2 93.4 93.5 |
| F1     | 75.6 85.1 75.0 77.9 | 87.8 92.6 89.7 91.3 | 86.6 87.4 86.7 87.8 |
Table D.3. Statistical estimators measured in each redshift bin for the EXP1a experiment.

| k-fold | Global | k-fold | Global | k-fold | Global | k-fold | Global |
|--------|--------|--------|--------|--------|--------|--------|--------|
| $z_{CLM} \in (0.18, 0.32)$ | 86.4 ± 1.1 | 86.2 | 89.0 ± 1.2 | 89.2 | 88.8 ± 1.4 | 88.6 | 88.1 ± 1.0 | 87.9 | 89.6 ± 1.3 | 89.6 |
| $z_{CLM} \in (0.32, 0.37)$ | 84.9 ± 2.6 | 84.1 | 87.0 ± 1.7 | 86.9 | 87.9 ± 1.6 | 87.3 | 87.1 ± 1.0 | 87.0 | 87.7 ± 2.0 | 87.3 |
| $z_{CLM} \in (0.37, 0.41)$ | 89.6 ± 1.6 | 89.2 | 92.1 ± 1.8 | 92.4 | 90.3 ± 1.5 | 90.3 | 89.5 ± 1.6 | 89.2 | 92.8 ± 0.9 | 92.7 |
| $z_{CLM} \in (0.41, 0.46)$ | 86.9 ± 0.9 | 86.6 | 89.3 ± 1.2 | 89.5 | 89.0 ± 1.3 | 88.8 | 88.2 ± 1.0 | 88.1 | 90.0 ± 1.2 | 89.9 |
| $z_{CLM} \in (0.46, 0.60)$ | 89.3 ± 1.1 | 88.5 | 91.7 ± 1.7 | 91.8 | 90.2 ± 1.5 | 89.9 | 89.4 ± 1.5 | 88.9 | 92.5 ± 0.9 | 92.2 |

Notes. Due to the k-fold approach, the performances are reported as pairs of mean and error (evaluated on the 10 folds) and as a single global value.

Fig. D.1. Comparison among the four band configurations (see Sect. 2), in terms of F1 score and average efficiency (AE) percentages (top panels), together with their square root of variances (bottom panels), as the number of spectroscopic sources in the training set increases (EXP1). In all panels, the linear best-fit trends are displayed as dashed lines. Due to the k-fold approach, performances have been averaged over the 10 folds, i.e. the x-axis shows the dimension of the training set, thus, the k-est fold used as test set has a size of $N/9$. 
**Fig. D.2.** Venn diagrams reporting the percentages of membership predictions performed by three different methods (CNN, RF, and Bayesian), measured on the common blind test set, obtained by combining the four clusters R2248, M0416, M1206, and M1149 (EXP3). On the columns, the common areas refer to the available shared sources, respectively, 460 CLMs and 519 NCLMs). On the rows, common predictions are split between correct and incorrect classifications. Global commonalities can be derived by summing values on the rows.

**Fig. D.3.** Comparison between the image-based CNN and two photometric catalogue-based approaches, RF and Bayesian method (EXP3) in terms of ROC curves for the four clusters: R2248 (top-left panel), M0416 (top-right panel), M1206 (bottom left panel), M1149 (bottom right panel).
Table D.4. Percentage performances on a blind test set related to the EXP2 experiment.

| Stacked  | Narrow | Intermediate | Large |
|----------|--------|--------------|-------|
| AE       | 84.5 ± 0.6 | 85.5 ± 0.4  | 86.6 ± 0.3 |
| CLM pur  | 79.6 ± 1.2  | 83.2 ± 0.2  | 82.5 ± 0.6 |
| % comp   | 87.6 ± 0.8  | 83.9 ± 0.8  | 88.5 ± 0.4 |
| F1       | 83.3 ± 1.2  | 83.6 ± 0.2  | 85.4 ± 0.6 |
| A370     |           |              |       |
| AE       | 85.4 ± 0.7  | 86.6 ± 0.3  | 87.4 ± 0.3 |
| CLM pur  | 80.3 ± 1.4  | 84.5 ± 0.2  | 83.9 ± 0.7 |
| % comp   | 86.5 ± 0.7  | 83.1 ± 0.6  | 86.6 ± 0.6 |
| F1       | 83.3 ± 1.4  | 83.8 ± 0.2  | 85.1 ± 0.7 |
| M0329    |           |              |       |
| AE       | 81.7 ± 0.5  | 83.5 ± 0.5  | 84.8 ± 0.3 |
| CLM pur  | 76.9 ± 0.7  | 79.2 ± 0.5  | 79.2 ± 0.4 |
| % comp   | 90.0 ± 0.6  | 90.4 ± 0.4  | 93.9 ± 0.4 |
| F1       | 82.9 ± 0.7  | 84.4 ± 0.5  | 85.9 ± 0.4 |
| MS2137   |           |              |       |
| AE       | 84.1 ± 1.1  | 82.4 ± 2.3  | 85.4 ± 0.7 |
| CLM pur  | 81.5 ± 1.9  | 84.4 ± 1.8  | 82.3 ± 1.0 |
| % comp   | 87.6 ± 1.5  | 77.1 ± 4.8  | 88.9 ± 0.7 |
| F1       | 84.2 ± 1.9  | 80.0 ± 1.8  | 85.4 ± 1.0 |

Notes. Performances have been split between the three test clusters: A370 (z = 0.375), MS2137 (z = 0.316), M0329 (z = 0.450) and their stacking. Best results are emphasised in bold. For ease of reading, only statistics related to the CLM class are reported, together with the average efficiency (AE), which refers to both classes.

Table D.5. Comparison between our image-based CNN model and two different photometric catalogue-based approaches, referred to the EXP3 experiment.

|                | CNN | RF | Bayesian | Δ   |
|----------------|-----|----|----------|-----|
| R2248 z = 0.346 |     |    |          |     |
| AE             | 88.1| 86.5| 85.9     | 1.6 |
| CLM compl      | 89.4| 87.0| 87.7     | 0.6 |
| F1             | 87.1| 87.5| 87.8     | 1.3 |
| NCLM compl     | 86.1| 85.1| 94.4     | -6.5|
| F1             | 87.0| 85.1| 94.4     | -6.5|

|                | CNN | RF | Bayesian | Δ   |
|----------------|-----|----|----------|-----|
| M0416 z = 0.397|     |    |          |     |
| AE             | 92.2| 89.2| 87.1     | 3.0 |
| CLM compl      | 87.1| 86.5| 91.2     | -4.1|
| F1             | 91.5| 89.7| 87.8     | 1.8 |
| NCLM compl     | 96.9| 92.3| 82.7     | 4.6 |
| F1             | 91.5| 88.3| 86.2     | 3.2 |

μΔ
-0.91 ± 1.42

|                | CNN | RF | Bayesian | Δ   |
|----------------|-----|----|----------|-----|
| M1206 z = 0.439|     |    |          |     |
| AE             | 89.7| 87.9| 85.0     | 1.8 |
| CLM compl      | 86.5| 81.9| 91.2     | -4.7|
| F1             | 88.2| 85.9| 85.3     | 2.3 |
| NCLM compl     | 92.3| 92.9| 79.4     | -0.6|
| F1             | 90.9| 89.7| 84.7     | 1.2 |

μΔ
-0.24 ± 0.90

|                | CNN | RF | Bayesian | Δ   |
|----------------|-----|----|----------|-----|
| M1419 z = 0.542|     |    |          |     |
| AE             | 89.4| 86.9| 85.5     | 2.5 |
| CLM compl      | 91.3| 88.5| 98.0     | -6.7|
| F1             | 86.6| 83.4| 82.9     | 3.2 |
| NCLM compl     | 88.3| 86.0| 78.4     | 2.3 |
| F1             | 91.3| 83.4| 87.4     | 3.9 |

μΔ
0.66 ± 1.60

Notes. The comparison involves two different model: a Random Forest and a Bayesian method, applied on photometric tabular information of four clusters: R2248 (z = 0.346), M0416 (z = 0.397), M1206 (z = 0.439) and M1419 (z = 0.542). Last column (Δ) shows the difference between CNN estimators and the best between the two photometric approaches, i.e. Δ = \text{estim}_\text{CNN} - \text{max}\{\text{estim}_{\text{pur}}, \text{estim}_{\text{AE}}\} for estim ∈ \{pur, compl, F1, AE\}, while rows μΔ list the averages among these Δ for each cluster.