



<b>Publication Year</b>	2019
<b>Acceptance in OA @INAF</b>	2020-12-10T17:06:00Z
<b>Title</b>	Astroinformatics-based search for globular clusters in the Fornax Deep Survey
<b>Authors</b>	Angora, G.; BRESCIA, Massimo; CAVUOTI, STEFANO; Paolillo, Maurizio; Longo, G.; et al.
<b>DOI</b>	10.1093/mnras/stz2801
<b>Handle</b>	<a href="http://hdl.handle.net/20.500.12386/28782">http://hdl.handle.net/20.500.12386/28782</a>
<b>Journal</b>	MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY
<b>Number</b>	490

# Astroinformatics-based search for globular clusters in the Fornax Deep Survey

G. Angora<sup>1,2★</sup>, M. Brescia<sup>1,2★</sup>, S. Cavuoti<sup>1,2,3,4★</sup>, M. Paolillo<sup>1,2,3,4</sup>, G. Longo<sup>3,4,5†</sup>,  
M. Cantiello<sup>6</sup>, M. Capaccioli<sup>2,3</sup>, R. D’Abrusco<sup>7</sup>, G. D’Ago<sup>1,8</sup>, M. Hilker<sup>9</sup>, E. Iodice<sup>2</sup>,  
S. Mieske<sup>10</sup>, N. Napolitano<sup>11</sup>, R. Peletier<sup>1,12</sup>, V. Pota<sup>2</sup>, T. Puzia<sup>1,8</sup>, G. Riccio<sup>1,2</sup>  
and M. Spavone<sup>1,2</sup>

<sup>1</sup>Department of Physics and Earth Science of the University of Ferrara, Via Saragat 1, I-44122 Ferrara, Italy

<sup>2</sup>INAF – Astronomical Observatory of Capodimonte, via Moiariello 16, I-80131 Napoli, Italy

<sup>3</sup>Department of Physics ‘E. Pancini’, University of Naples Federico II, via Cinthia 21, I-80126 Napoli, Italy

<sup>4</sup>INFN – Napoli Unit, via Cinthia 21, I-80126 Napoli, Italy

<sup>5</sup>Department of Astronomy, California Institute of Technology, Pasadena, CA 90125, USA

<sup>6</sup>INAF – Astronomical Observatory of Abruzzo, Via Mentore Maggini snc, Loc. Collurania, I-64100 Teramo, Italy

<sup>7</sup>Center for Astrophysics | Harvard & Smithsonian, 60 Garden St, Cambridge, MA 02138, USA

<sup>8</sup>Institute of Astrophysics, Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860, 7820436 Macul, Santiago, Chile

<sup>9</sup>European Southern Observatory, Karl-Schwarzschild-Str 2, D-85748 Garching, Germany

<sup>10</sup>European Southern Observatory, Alonso de Cordova 3107, 7630355 Vitacura, Santiago, Chile

<sup>11</sup>School of Physics and Astronomy, Sun Yat-sen University, Zhuhai Campus, Guangzhou 519082, P.R. China

<sup>12</sup>Kapteyn Astronomical Institute, University of Groningen, PO Box 800, NL-9700 AV Groningen, the Netherlands

Accepted 2019 September 30. Received 2019 September 28; in original form 2019 May 7

## ABSTRACT

In the last years, Astroinformatics has become a well-defined paradigm for many fields of Astronomy. In this work, we demonstrate the potential of a multidisciplinary approach to identify globular clusters (GCs) in the Fornax cluster of galaxies taking advantage of multiband photometry produced by the VLT Survey Telescope using automatic self-adaptive methodologies. The data analysed in this work consist of deep, multiband, partially overlapping images centred on the core of the Fornax cluster. In this work, we use a Neural Gas model, a pure clustering machine learning methodology, to approach the GC detection, while a novel feature selection method ( $\Phi$ LAB) is exploited to perform the parameter space analysis and optimization. We demonstrate that the use of an Astroinformatics-based methodology is able to provide GC samples that are comparable, in terms of purity and completeness with those obtained using single-band *HST* data and two approaches based, respectively, on a morpho-photometric and a Principal Component Analysis using the same data discussed in this work.

**Key words:** methods: data analysis – globular clusters: general – galaxies: elliptical and lenticular, cD.

## 1 INTRODUCTION

In modern observational Astronomy, the amount of data collected by an instrument in a single day is often more than enough to keep occupied an entire community of scientists for long time; LSST, for instance, will produce 20 trillion bytes of raw data per night.<sup>1</sup> These huge data sets are further enlarged by the possibility

to combine data obtained at different wavelengths and epochs by different instruments. Astronomy, in fact, is going to enter the big data era not only for the sheer size of its data, but also for the high dimensionality and complexity of the parameter spaces (PSs) to be explored. These spaces are composed by a variable mixture of photometry, spectroscopy, structural and morphological features, depending on the specific context of the problem under investigation. A complexity that allows, on the one hand, to answer long-standing questions with higher accuracy and, on the other hand, to address completely new and more difficult problems. In such scenario a new paradigm is required, mostly based on a multidisciplinary approach, through a virtuous integration of Astrophysics, Data Science, and Informatics. A symbiosis that is at

\* E-mail: [gius.angora@gmail.com](mailto:gius.angora@gmail.com) (GA); [massimo.brescia@inaf.it](mailto:massimo.brescia@inaf.it) (MB); [stefano.cavuoti@gmail.com](mailto:stefano.cavuoti@gmail.com) (SC)

† Merle Kingsley distinguished visitor.

<sup>1</sup><https://www.lsst.org/about/dm>

the very heart of the relatively new discipline of Astroinformatics, or Knowledge Discovery in astrophysical data (Borne et al. 2009; Brescia et al. 2013; Feigelson & Hilbe 2014; Brescia et al. 2018). Astroinformatics, however, is just a label which summarizes the emerging awareness that complex problems can be tackled only by heterogeneous groups of experts, and that multidisciplinary approach is not a presumptuous ambition, but rather an unavoidable and precious quality. In the last decade, in many different fields, it has been clearly demonstrated that the emulation of the mechanisms underlying natural intelligence, if translated into efficient algorithms and supplied to super computers, is fully and rapidly able to analyse, correlate, and extract huge amounts of heterogeneous information (Brescia et al. 2018; Baron 2019).

When dealing with high dimensionality PSs, it appears evident the crucial importance of an optimal choice of the PS (i.e. feature selection, hereafter FS) adopted to represent the data to be explored in the context of a specific problem.

The selection of an optimal set of features strictly depends on the concept of *feature importance*, based on the quantification of its *relevance*.

Formally, the importance of a feature is the relevance of its informative contribution to the solution of a learning problem. There is plenty of FS solutions proposed in literature (Guyon & Elisseeff 2003), such as Principal Component Analysis (PCA; Jolliffe 2002), *filter* techniques (Gheyas & Smith 2010), *wrapper* (Kohavi & John 1997), and *embedded* methods (Lal et al. 2006), among which a typical example is the Random Forest (RF) model (Breiman 2001). These methods are basically oriented to find the smallest (best) PS able to solve a given problem (Jain & Zongker 1997; Guyon et al. 2006; Hastie, Tibshirani & Friedman 2009).

Such multidisciplinary paradigms have been concretely followed in this work by exploiting automatic self-adaptive methodologies (e.g. FS and the two machine learning (ML) paradigms, respectively, supervised and unsupervised learning; Russell & Norvig 2010) with the main goal of identifying globular clusters (hereafter GCs) in the Fornax cluster of galaxies.

GCs represent an important category of widely studied astronomical sources. Since GCs harbour a wide variety of stellar types of the same age, each single GC acts as a stellar laboratory, suitable to observe and analyse the formation, behaviour, and evolution of stellar systems concentrated within just  $\sim 10$  parsec. As a population, on a galactic scale, they trace the dynamics, the kinematics, and the chemistry of their host galaxy, behaving like a sort of a footprint left by the galactic evolution (Brodie & Strader 2006; Ashman & Zepf 2008).

It is now well established that GC can be split in populations (e.g. Geisler & Forte 1990; Zepf, Ashman & Geisler 1995; Brodie & Strader 2006; Pota et al. 2013): (i) a red, metal-rich, spatially concentrated sub-population and (ii) a blue, metal-poor, spatially extended sub-population.

The data analysed in this work consist of deep, multiband images, partially overlapped, centred on the core of the Fornax cluster. The extracted catalogue is composed by several thousands of sources, each one characterized by a large set of features (i.e. parameters), such as luminosity, colours, and morphological information, for a total of more than 60 features. Given the high number of dimensions involved, the difficulty to disentangle different types of objects (e.g. foreground stars, background galaxies, and GCs), together with the fact that spectroscopic confirmation was available only for a quite limited number of sources, it was decided to tackle the task of recognizing and classifying GCs (against a variety of background and foreground sources) by investigating both the

PS optimization and the classification capabilities of specific ML methods.

The work has therefore focused on the Growing Neural Gas model (GNG; Martinetz & Schulten 1991; Martinetz, Berkovich & Schulten 1993; Fritzke 1994), a pure clustering category of ML methods, together with a novel FS method, named  $\Phi$ LAB (Brescia et al. 2019). In order to compare the performance of the Neural Gas-based model, a Multi-Layer Perceptron with Quasi-Newton approach (MLPQNA; Byrd, Nocedal & Schnabel 1994; Bortoletti et al. 2003; Brescia et al. 2012; Cavuoti et al. 2012) and a  $K$ -means (Bishop 2006) have been used as a test benchmark. While, in order to evaluate the FS performances, GNG and RF (Breiman 2001) methods have been compared on several data sets. To evaluate the accuracy and in particular the efficiency in identifying secure GC candidates, a direct comparison of these methods has been performed with other techniques as well as with very promising results obtained with other types of ML methods applied on single-band *HST* data of NGC 1399, the giant elliptical galaxy at the centre of the Fornax cluster. An important corollary aspect of this work was, in fact, to evaluate the level of accuracy in GC classification within two different contexts: multiband ground-based data and the single-band high spatial resolution data obtained from space.

The paper is structured as follows: In Section 2, we describe the data used in this work. In Section 3, we present the methods employed for the experiments. In Section 4, we describe and discuss the experiment results. Section 5 is dedicated to a comparison with similar ML experiments performed on *HST* data and with other approaches. In Section 6, we estimate the density maps of the GCs spatial distribution as further validation method. Finally, in Section 7, we draw our conclusions.

## 2 DATA

The data used in this work cover the central region of the Fornax cluster and were obtained with the OmegaCam (Kuijken 2011) camera, installed on the VLT Survey Telescope (VST; Schipani et al. 2012) as part of the Fornax Deep Survey (FDS; Iodice et al. 2016, Peletier et al., in preparation).<sup>2</sup> The images were obtained through 76 exposures of 150 s in the  $u$  band, 54 s in the  $g$  and  $r$  bands, and 35 s in the  $i$  band, reaching a  $S/N \sim 10$  at, respectively, 23.8, 24.8, 24.3, 23.3 magnitudes in the  $u$ ,  $g$ ,  $r$ ,  $i$  (D’Abrusco et al. 2016). The average seeing was  $1.17 \text{ arcsec} \pm 0.08 \text{ arcsec}$  in the  $g$  band and  $0.87 \text{ arcsec} \pm 0.07 \text{ arcsec}$  in the  $r$  band ( $u$  and  $i$  bands show similar variations over the observed field).

The catalogue,<sup>3</sup> extracted using SEXTRACTOR (Bertin & Arnouts 1996), consists of 94 067 sources whose right ascension (RA) and declination (Dec.) are inside the celestial square of limits  $\sim [54.02, 55.38] \times [-34.91, -36.03]$  (measured in degrees). The catalogue does not contain the same number of sources in each band, due to the different depth of observations in different filters: there are 15 095 sources in the  $u$  band, 73 497 sources in  $g$  band, 72 385 in  $r$  band, and 49 207 in  $i$  band. For each band and for each source SEXTRACTOR

<sup>2</sup>FDS is an ESO joint program, based on two Guaranteed Time Observation surveys, FOCUS (P.I. R. Peletier) and VEGAS (P.I. E. Iodice; Capaccioli et al. 2015), having as main goal the study of the whole Fornax cluster out to its viral radius.

<sup>3</sup>The catalogues for the full FDS survey will be presented in a forthcoming paper (Cantiello et al., in preparation). Here, we adopted the catalogues used in D’Abrusco et al. (2016) and Cantiello et al. (2018b).

was used to derive the following information (Bertin & Arnouts 1996):

- (i) the automatic aperture magnitudes with error (*MAG\_AUTO*), i.e. an estimation of the total magnitude;
- (ii) the fixed aperture magnitudes (*MAG\_APER*): an estimation of the flux above the background within different circular apertures (4, 6, 8, 16, and 32 pixels, respectively), with the related errors;
- (iii) the peak surface brightness above background (*MU\_MAX*);
- (iv) the average full width at half-maximum (FWHM) of the image assuming a Gaussian core (*FWHM\_IMAGE*). It is the average, due to the various overlaps, by considering the small variations among the fields;
- (v) the semimajor and semiminor axes lengths (*A\_WORLD*, *B\_WORLD*) with the errors;
- (vi) the position angle between the major axis and the  $x$ -axis of the image (*THETA\_WORLD*);
- (vii) the ratio between the semimajor and semiminor axes lengths (*ELONGATION*);
- (viii) the fraction-of-light radii. It measures the radius of the circle centred on the barycentre that encloses about half of the total flux (*FLUX\_RADIUS*);
- (ix) Kron apertures (*KRON\_RADIUS*), within  $2.5 \times FLUX_RADIUS$ ;
- (x) The Petrosian apertures (*PETRO\_RADIUS*), i.e. the apertures defined by the Petrosian radius, i.e. the radius limit of the ratio between the local surface brightness and the mean interior surface brightness of the source.

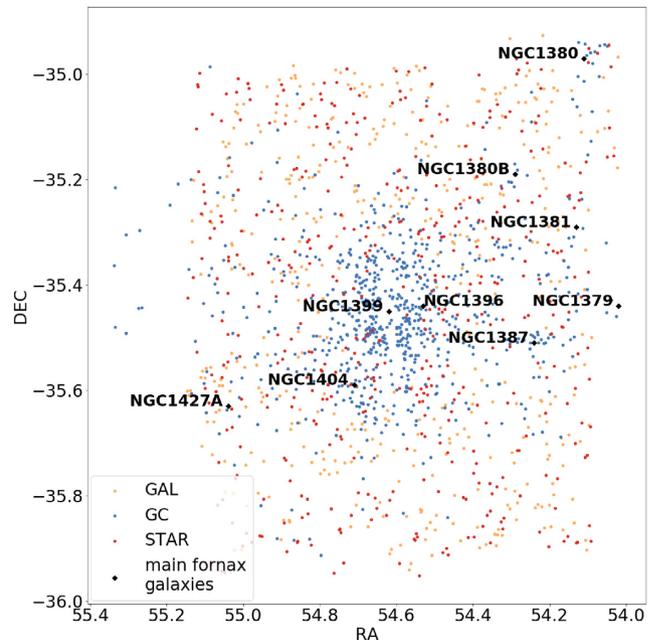
By adding colours ( $u-g$ ,  $g-r$ , and  $r-i$ ) and by excluding the two larger apertures to minimize contamination from nearby sources and to limit the magnitude errors induced by the background contamination, the final PS consists of 64 features: 16 magnitudes, 36 photometric parameters, and 12 colours.

To build the knowledge base (KB) needed for the training of the ML network, we used a set of spectroscopically confirmed sources obtained by combining the catalogues from Pota et al. (2018), consisting of newly confirmed GCs and previous data sets from Wittmann et al. (2016), which is mostly based on Schubert et al. (2010). In addition, the foreground stars were provided by Pota et al. (2018) and the background galaxies by D'Ago et al. (in preparation). The sky distribution of the various objects is illustrated in Fig. 1 for both the spectroscopically confirmed objects and the unclassified sources.

Since these catalogues were derived with different instruments and methodologies, we applied our in-house cross-matching method (Riccio et al. 2017) between GCs and galaxies and the FDS catalogue, imposing a matching tolerance of 0.25 arcsec. Stars were not cross-matched because they are already available in the FDS catalogue.

After the cross-matching procedure, 1627 sources were labelled: 706 GCs, 464 foreground stars, and 457 background galaxies. However, not all these labelled objects turned out to be suitable to construct the KB, due to the presence of missing data (i.e. missing values in some feature columns). In particular, the missing data for the 1627 cross-matched sources were ( $u$ ) 509 (31.3 per cent), ( $g$ ) 6 (0.4 per cent), ( $r$ ) 8 (0.5 per cent), ( $i$ ) 5 (0.3 per cent). While the missing data for the whole catalogue (94 067 sources) were ( $u$ ) 78 971 (84.0 per cent), ( $g$ ) 20 531 (21.8 per cent), ( $r$ ) 21 637 (23.0 per cent), ( $i$ ) 44 822 (47.6 per cent).

Most of these missing data are in the  $u$  band (i.e. the less deep). However, in this case, missing information is mostly due to the sen-



**Figure 1.** Distribution of spectroscopically sources: GCs (blue), foreground stars (red), background galaxies (yellow), and bright Fornax cluster galaxies (black diamonds).

sibility limit of the instrument, rather than to the presence of *holes* in the data distributions, causing an intrinsic difficulty to test any imputation method. In fact, although there are numerous imputation techniques (Poulos & Valle 2016; Yoon, Jordon & van der Schaar 2018; Zhang, Xie & Xing 2018; Camino, Hammerschmidt & State 2019), able to predict missing values within the sample features, the prediction of feature values outside the training distribution is a more tricky and complex problem, beyond the goals of this work. We excluded all of them from the final sample due to the known negative impact of such missing information on the performances of ML models (Batista & Monard 2003; Marlin 2008; Parker 2010; Brescia et al. 2019). We did not introduce any further error-based cuts in order to avoid any additional reduction of the KB.

### 3 THE METHODS

In this work, we make use of an optimized implementation of the GNG network (Fritzke 1995) obtained using the Theano programming environment (The Theano Development Team et al. 2016), and a novel FS method,  $\Phi$ LAB, to optimize the PS. Moreover, we briefly introduce the three methods used as test benchmark. Such models are described in the following sections. For all these networks, the hyperparameters have been set by following a heuristic pruning process.

#### 3.1 Growing Neural Gas

The GNG model was introduced in Fritzke (1994) as a variant of the Neural Gas algorithm (Martinetz & Schulten 1991), which combines the Competitive Hebbian Learning (Martinetz et al. 1993) with a vector quantization technique, to achieve a learning that retains the topology of the data set. This is an important property, since the vector quantization introduces an order relationship between the data PS and the internal architecture of the network.

In fact, Vector quantization techniques (Martinetz et al. 1993), encode a data manifold, e.g.  $V \subseteq R^m$ , using a finite set of reference vectors  $w = w_1 \dots w_N$ ,  $w_i \in R^m$ ,  $i = 1 \dots N$ . Every data vector  $v \in V$  is described by the best matching unit (BMU), i.e. the neural unit whose reference vector  $w_{i(v)}$  minimizes the distortion error  $d(v, w_{i(v)})$ . This procedure divides the manifold  $V$  into a number of subregions:  $V_i = \{v \in V : \|v - w_i\| \leq \|v - w_j\| \forall j\}$ , called Voronoi polyhedra (Montoro & Abascal 1993), within which each data vector  $v$  is described by the corresponding reference vector  $w_i$ . The BMU and the second-BMU develop a connection that, if not energized again during learning, tends to decay and then to be removed (Fritzke 1994). The GNG network is characterized by a variable number of neurons during the learning phase: new units are added to an initially small number of units through the estimation of a statistical local measure obtained during the previous adaptation steps, while isolated units are removed. The insertion mechanism has to be able to find the location in the PS where to introduce a new neuron, in order to reduce the reconstruction error. In other words, the insertion mechanism finds subregions of the data manifold whose reconstruction is more complex, i.e. the subregions characterized by a relatively high density.

Each neuron has an attribute defined as the local error  $E_i$ , whose value is updated at each iteration only for the BMU  $i_0$ :  $\Delta E_{i_0} = \|w_{i_0} - v\|^2$ , where  $v$  is the extracted input vector. After a certain number of iterations,  $E_i$  represents a local reconstruction error for the neural unit  $i$ . Units characterized by high values of  $E_i$  are associated with large Voronoi polyhedra, and these regions require better sampling to be correctly reconstructed.

The adaptation rule is applied only for the BMU  $i_0$  and for its topological neighbours:

$$\Delta w_{i_0} = \epsilon_w \cdot (v - w_{i_0}),$$

$$\Delta w_j = \epsilon_n \cdot (v - w_j) \quad \forall j \in \text{Neighbours}(i_0). \quad (1)$$

The advantage of this method is that the learning is completely determined by the input data, i.e. it is not necessary to superimpose a structure to the network as, for instance, the expected number of clusters. The downside is the single input extraction at each iteration, which leads to an extra computational cost on large data set. For this reason we optimized the GNG implementation using Theano (The Theano Development Team et al. 2016), an open source PYTHON library allowing an efficient computation of tensor mathematical expressions and an easy exploitation of the Graphical Processing Unit (GPU). Furthermore, we revised the adaptation rule of equation (1), by introducing a gradient descent method with respect to the cost function, represented by the quantization error:

$$\Delta W = -\eta \nabla_W(Q\mathcal{E}),$$

$$Q\mathcal{E} = \frac{1}{2|V|} \sum_{i=1}^p \sum_{n \in \text{BMU}_i} \|v_n - w_i\|^2, \quad (2)$$

where we have assumed that:  $v_n$  is the  $n$ th input vector mapped by the BMU  $i$  whose reference vector is  $w_i$ ,  $V$  is the data manifold composed by  $|V|$  records,  $p$  is the number of BMUs. Finally, we added also a batch extraction criteria, i.e. at each iteration a subset of sources has been extracted from the data, whose dimension is between 1 (equivalent to the original case) and the full data set.

### 3.2 $\Phi$ PhiLAB: a novel feature selection method

We recently investigated the possibility to find a PS optimization method able to cope with the all-relevant FS requirements and to

infer knowledge within the data-driven analysis domain, hence particularly suitable for astrophysical problems. The designed method is called  $\Phi$ LAB (PhiLAB, Parameter handling investigation LABORatory), which aims at identifying the exact PS, by solving the so-called *all-relevant* FS problem (Brescia et al. 2019; Delli Veneri et al. 2019). The method is a hybrid approach, including properties of both wrappers and embedded FS categories (Tangaro et al. 2015). It is, in fact, based on two joined concepts: *shadow features* (Kursa & Rudnicki 2010) and *Naïve LASSO* statistics (Least Absolute Shrinkage and Selection, (Hastie, Tibshirani & Friedman 2001; Tibshirani 2013; Hara & Maehara 2016; Hara & Maehara 2017), by using the RF (Breiman, Last & Rice 2003) as feature importance computing engine. By joining the two concepts,  $\Phi$ LAB is able to determine a threshold to filter the most relevant candidate features and to refine the final selection by determining the additional weak relevant features through an  $L_1 - norm$  regularization of a ridge regression (Tikhonov 1998), retaining only the non-zero features representing the optimal solution.

### 3.3 Benchmark methods

In order to compare the classification capability of the GNG and to explore the features selection performed by  $\Phi$ LAB, we used three methods, described in the following. Having the possibility to use both supervised and unsupervised models, we tried also to perform a comparison between the two categories, by taking into account that, although supervised paradigm is generally preferred whenever a KB is available, we were particularly interested to evaluate the performances of the GNG model in a complex astrophysical problem.

(i) MLPQNA: It is a very robust supervised ML model, as it has been already demonstrated by its capability to achieve high performances on a variety of astrophysical problems (Cavuoti et al. 2012, 2015; Brescia et al. 2012; Brescia & Longo 2013). For this reason we have chosen this model as upper limit benchmark method. Its architecture is similar to an MLP (Bishop 2006), with a Quasi-Newton algorithm used as optimizer furthermore it makes use of the known L-BFGS algorithm (Limited memory; Broyden Fletcher Goldfarb Shanno, Byrd et al. 1994). This network has been applied as test benchmark of GNG performances (results shown in Section 4.2). The network is composed by two hidden layers, respectively with  $2N + 1$  and  $N - 1$  neural units, where  $N$  is the number of input dimensions (i.e. the number of features). The neuron activation function is a hyperbolic tangent. Furthermore, the network weight updating is based on the L2-norm regularization term (Bishop 2006), with a decay factor of 0.01.

(ii) RF: It is a widely known supervised ML ensemble method that uses a random subset of features to build a collection of decision trees (Breiman 2001). The method is characterized by an intrinsic absence of training overfitting (i.e. excess of training data fitting with consequent poor fitting of blind test data). RF has been applied in order to verify the sensitivity of GNG to noisy and redundant data parameters and to investigate regarding the efficiency of  $\Phi$ Lab to individuate the best set of features. The results are shown in Appendix A and discussed in Section 4.1. The method has been trained with 500 trees, the gini index is used in order to evaluate the quality of the split (Breiman 2001), while the maximum number of features, required to search the best split, coincides with the involved number of parameters. Furthermore, the minimum number of samples necessary to split a node has been set equal to 2 and no limit was imposed to the depth of the tree growth.

(iii) *K*-means: It is a clustering method able to partition the data set into *K* clusters, minimizing the distortion measure. At the end of the training phase the data set has been divided into *K* Voronoi polyhedra (Montoro & Abascal 1993). Such method provides a benchmark lower limit: although the model is able to perform a vector quantization process, it is necessary to overimpose a structure on the data, i.e. the number of cluster is not automatically determined. Thus, in order to compare the unsupervised networks, we trained the model by setting *K* to the number of clusters found by the GNG. Finally, the expectation maximization (EM) algorithm was used to train the model.

## 4 EXPERIMENTS

We generated two different data sets to train our models, one including the *u* band and the second excluding it. This partition was imposed by the fact that in FDS the *u* band is much shallower than the others (among the  $\sim 94\,000$  sources in the catalogue, only 15 per cent of the objects were detected in *u* band, against higher percentages for *g*, *r*, and *i* bands, respectively, 78 per cent, 77 per cent, and 53 per cent). The resulting data sets can be summarized as it follows: (a) *ugri* data set with 1113 objects, of which 357 are GCs, 416 galaxies, and 340 stars; (b) *gri* data set with 1618 objects, including 699 GCs, 457 galaxies, and 462 stars. For both data sets we performed three different classification experiments:

- (i) a 3-class problem, i.e. stars, GCs, and galaxies (named as *3CLASS*);
- (ii) a 2-class problem by grouping in the same class stars and galaxies against GCs (named as *GCALL*);
- (iii) a 2-class problem, namely stars versus GCs (named as *GCSTAR*).

Hence, we performed a total of six experiments. Due to the limited amount of labelled samples available (i.e. sources with a known spectroscopic classification), a canonical splitting of the KB into training ( $\sim 80$  per cent) and blind test set ( $\sim 20$  per cent) could not be applied. In order to circumvent this problem and to balance the samples for each class during the learning phase, the training-test experiments involved an approach based on the stratified *k*-fold cross-validation (Hastie et al. 2009; Kohavi 1995): the KB is split into five non-overlapping subsets. In this way, by iteratively taking each time four of these subsets as training set, and using the fifth as blind test set, an overall blind test on the entire KB available can be performed. As it was already mentioned, in order to perform the optimal choice of the PS, we applied the method  $\Phi$ LAB, by identifying a proper subset of features for each of the six experiments. We use an RF method, together with the GNG, to analyse the selection achieved by  $\Phi$ LAB: we measured the model performances by varying the PS (see the next section).

The statistical estimators used to measure both the FS and classification performances are (i) the average efficiency (AE), the ratio between the number of correctly classified objects and the total number of objects, averaged over all involved classes; (ii) the purity (*pur*, also called *precision*), i.e. the ratio between the number of correctly classified objects of a class and the total number of objects classified in that class (the dual of the contamination); (iii) the completeness (*compl*, also called *recall*), i.e. the ratio between the number of correctly classified objects of a class and the total amount of objects of that class; (iv) the F1

score (*F1*), defined as two times the ratio between the product of purity and completeness and their sum, for each class (Stehman 1997).

### 4.1 Feature selection

In order to analyse and optimize the PS suitable for classification purposes, the six data sets were first processed by  $\Phi$ LAB. In this work, our main interest was to obtain the most simplified PS able to predict a good classification on new objects (i.e. outside the data used for training + test) based on the training on the full data set available. That is why, along the reported experiments, we took in consideration the results of FS applied always to the whole data set. As an example, the feature importance related to the *ugri-GCALL* case is shown in Fig. 2, where the PS is partitioned into a *rejected* set of features (considered to be irrelevant in terms of information contribution) and a retained set of *all relevant* features, composed by both *best* and *weakly relevant* features.

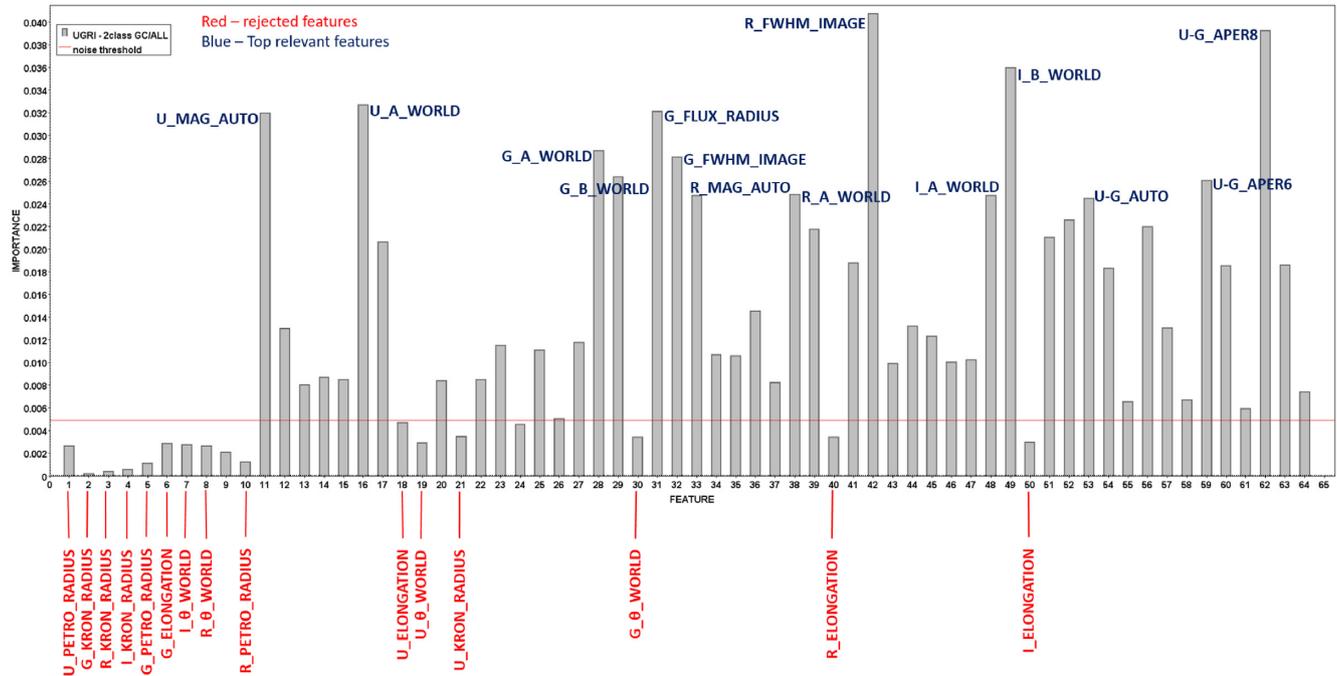
The results of the six FS experiments are summarized in Table A1 (Appendix A) for both the selected and the rejected subsets of features. By analysing such results, the *FWHM*, *FLUX\_RADIUS*, and colours appear relevant in all the six classification experiments (these three sets of features, with respect to the involved set of relevant features, have an informative contribution ranging in 9–31 per cent, 4–24 per cent, and 12–26 per cent, respectively), thus confirming the higher relevance of the colours over magnitudes, as well as the intrinsic importance of *FWHM* and *FLUX\_RADIUS*, indispensable to disentangle the extended objects from foreground point-like sources and the unresolved GCs. Concerning the *FWHM*, its informative contribution remains high, although being averaged to take into account small variations among the FDS fields. The features *MU\_MAX*, achieving a large informative contribution (ranging in 5–18 per cent), mark a slight difference between the experiments that include galaxies and those in which only the GC/star separation is required.

Moving from the *ugri* type to *gri* and from the *3CLASS* experiments to *GCSTAR*, there is a flattening of the informative contribution among the features, mainly due to the exclusion of the *u* band and by grouping together stars and galaxies, thus increasing the complexity of the classification problem.

Regarding the removed parameters: *THETA* features have a negligible informative contribution; *KRON\_RADIUS* carries a very weak contribution (the sum over all the involved filters does not reach 1 per cent of the whole informative contribution). The improvement due to the *ELONGATION* features is always much smaller than the information shared by the *SEMI-AXIS* features (sometimes about one order of magnitude). Although it may appear as a bit surprising that the elongation does not show a particular relevance, this can be due to the information already carried by the semi-axes that are an absolute quantity depending on the shape and distance of the object and its extension. In particular, the combination of the semi-axes and extension (e.g. *FLUX\_RADIUS*) may embed the elongation information.

The information carried by the *PETRO\_RADIUS* oscillates among the experiments and its informative contribution disuniformity is mostly due to *i* band. Thus, according to the *all-relevant FS* approach, this feature has not been rejected. From all these considerations, the resulting optimized PS extracted for the six classification experiments consists of 49 features (listed in upper Table A1 in Appendix A).

In order to verify that the PS extracted by  $\Phi$ LAB, is the best suitable set of representative features for the GC classification, we



**Figure 2.** Example of FS results for the experiment 2-class *ugri* (GCs versus stars + galaxies). The red line is the *shadow feature* noise threshold, defining the separation between *best* and *weakly relevant* features. The rejected features are in red, while the top relevant features are in blue. Table A1, in Appendix A, reports the feature importance values estimated by  $\Phi$ LAB for all the six experiments. Feature 9 (*i*-band *PETRO\_RADIUS*) and feature 24 (*g*-band *MAG\_APER6*), although under the shadow feature noise threshold, have not been considered as rejected, because retained as weak relevant after the application of the LASSO statistics (see Section 3.2 for details).

performed a test, based on the following training + test classification experiments, involving all six data sets:

- (i) *BEST PS*: using the optimized PS, composed by the 49 features extracted by the FS method, hence representing the best solution to the *all-relevant* FS problem;
- (ii) *FULL PS*: by using the full PS, composed by all 64 features available;
- (iii) *MIXED PS*: by altering the BEST PS, replacing a group of 15 randomly selected features with the 15 rejected features;
- (iv) *BEST + REJECTED PS*: by replacing the 15 least relevant features of the BEST PS with the 15 rejected features.

In all these tests, for each of the six data sets involved, the same training and test sets have been used, as well as the same configuration setup for the model GNG, in order to avoid any spurious effect on the classification statistics induced by the change of internal model parameters and by the data used for training and test. The resulting PSs are summarized in Table A2, while the related GNG performances are reported in Table A4, showing that

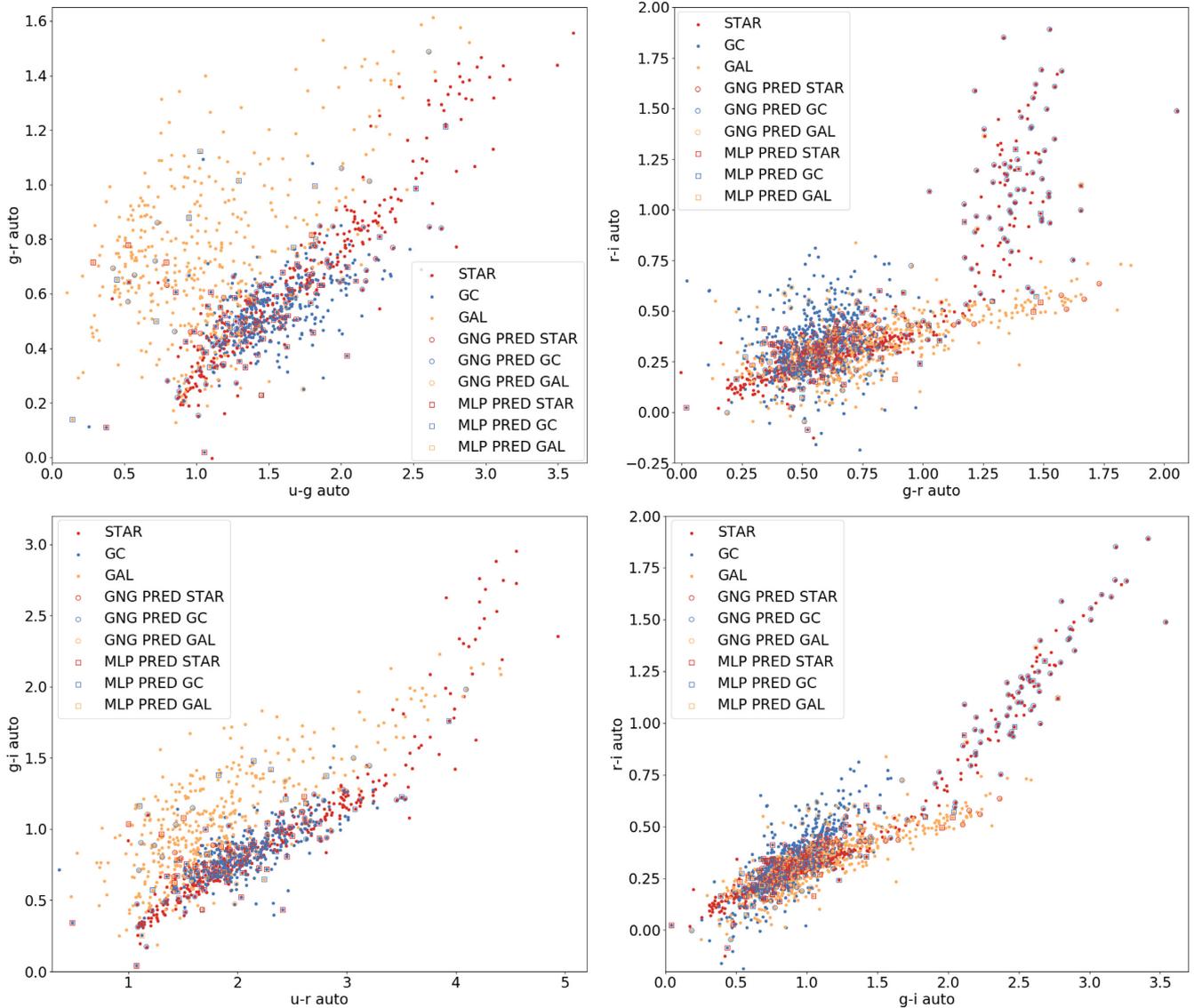
- (i) the selected set of features (named as BEST) allows the GNG to achieve high performances, reaching an increase of 50 per cent in terms of AE, i.e. the GNG trained on the BEST data set always reaches better scores in terms of statistical estimators, whereas the performance degradation is due only to the removal of the *u* band ( $\sim 8$  per cent in terms of AE), although always remaining well above the other PSs;
- (ii) the separation between GCs and notGCs (star and galaxies) appears to be the least complex problem for the GNG (showing an increase of the AE between  $\sim 3$  per cent and  $\sim 45$  per cent);
- (iii) concerning the other PSs (FULL, MIXED, BEST + REJECTED), the additional information carried by a greater number

of sources (*gri*) predominates on the information represented by the *u* band (*ugri*) only in the *GCALL* experiments (separation between GCs and *notGCs*), gaining up to 10 per cent in terms of AE. This trend is reversed when the classification involves the separation of the stars and galaxies (*3CLASS* and *GCs versus stars*): in these cases there is a greater dependence on the absence of the *u* band rather than on the number of sources (with differences of AE  $\lesssim 9$  per cent). The only exception occurs in the most complete and complex case, i.e. the *ugri 3CLASS* with FULL PS, where the amount of objects has a greater impact (with  $\lesssim 7$  per cent gain in terms of AE).

This analysis seems to support the idea that GNG is particularly sensitive to noisy or redundant features and, in order to verify this hypothesis, we repeated the same training/test process (i.e. the same data set for the same experiments) replacing the GNG with an RF method (Breiman 2001).

The RF method, by generating a random ensemble of decision trees, is very robust to parameter variations. The results are shown in A6, from which

- (i) in all cases, RFs trained on the data set BEST achieve the highest scores, showing an increasing of  $\lesssim 8$  per cent in terms of AE;
- (ii) concerning the *3CLASS* and *GCALL* problem, the *u* band improves the AE ( $\lesssim 8$  per cent) but, without this additional photometric information, BEST is always more robust than others ( $\lesssim 6$  per cent);
- (iii) the MIXED data set is always the worst PS, with a decrease larger than 5 per cent in terms of AE;
- (iv) in the *GCSTAR* case, the *u* band seems to lose partially its positive role (with an AE decreasing of  $\sim 2$  per cent) and, without



**Figure 3.** Misclassified sources plotted on colour–colour diagrams related to the *ugri* data set (left top panel,  $g-r$  versus  $u-g$  and left bottom  $g-i$  versus  $u-r$ ) and related to the *gri* data set (right top panel,  $r-i$  versus  $g-r$ , right bottom panel,  $r-i$  versus  $g-i$ ), together with the spectroscopic set. In all figures train GCs are plotted with blue dots, train stars with red dots, and train galaxies with orange dots. The incorrect predictions made by GNG are plotted with open circles, while MLPQNA misclassified sources are plotted with open squares. For both of them the incorrect classifications are coloured in red, blue, and orange, respectively, for sources predicted as stars, GCs, and galaxies.

the  $u$  band, the data set *FULL* is quite often more robust than the others ( $\lesssim 3$  per cent);

(v) Purity is almost stable in all PSs, while completeness presents large fluctuations (the largest purity deviation is  $\sim 2$  per cent, whereas the largest completeness is  $\sim 7$  per cent).

(vi) it appears that the absence of galaxies in the sample (i.e. *GCSTAR* cases) makes the experiments less sensitive to the presence of the  $u$ -band data (showing a decrease of AE of 0.5 per cent, i.e. the results show fluctuations around a mean value). While the presence of the  $u$  band affects more significantly the classification capability in presence of galaxies, making the *BEST* case the most powerful data set. The role of the  $u$  band can be derived from Figs 3 and 5, according to what was widely discussed in Muñoz et al. (2014), where the authors showed (see, for instance, their figs 13 and 16) how the bluer spectral energy distribution of star-forming galaxies and passive galaxies at moderate redshift is

well identified in the colour–colour diagrams. The  $u$  band, in this case based only on optical colours, shows much less discriminant power for GCs and stars, becoming more relevant in identifying galaxies.

Given these premises, in order to explore the impact of the  $u$  band on the classification, we performed two further experiments: (i) a data set extracted from the *ugri* set without the  $u$  band, named as *gri\** and (ii) by using only the features related to the  $u$  band. Such test has the role to disentangle the  $u$ -band contribution from the effect carried by the increase of the samples. Results are shown in Table A5 (in Appendix A), from which we conclude that

(i) the use of the single  $u$ -band information still allows the separation among classes with a slight decrease of AE for both GNG ( $\lesssim 2$  per cent) and RF ( $\sim 3$  per cent);

(ii) concerning the *gri* data set the average efficiency reduction appears higher if compared to the *ugri* data set ( $\gtrsim 5$  per cent);

(iii) RF performances significantly decrease with respect to the results achieved in the case where all the available samples are used (rising up an AE difference of  $\sim 10$  per cent in the same cases), showing the known dependence of ML methods from the training dimension (Brescia et al. 2013).

Hence, the experiments performed by the two models confirm:

(i) the strong dependence of GNG on the structure of the PSs and thus support the need for a robust method of features selection, (ii) the capability of  $\Phi$ LAB to identify the set of relevant features, (iii) the high impact of the *u* band on the classification capability for both models.

## 4.2 GC classification

As introduced in Section 3.3 the GC classification experiments have been performed by comparing the GNG model with MLPQNA and *K*-means. For the latter model the performances have been reported in Table A3 and represent a lower limit for the GNG. In terms of average efficiency, the GNG shows better results, gaining from 1.0 per cent to 6.6 per cent. For this reason we focused more on the comparison between MLPQNA and GNG. The classification results on the blind test sets, obtained by GNG and MLPQNA using the *k*-fold technique for the six data sets, are summarized in Table 1. In Fig. A1, the ROC curves (receiver operating characteristic; Hanley & McNeil 1982) have been reported to study the purity-completeness trade-off. In order to compare the results, it is important to remark that the unsupervised model (GNG) does not take into account the knowledge of the source labels during the learning, entrusting the weights adaptation to the minimization of the quantization error, while the supervised model (MLPQNA) uses labels to guide learning, allowing also the identification of the minimum with a very high efficiency (Brescia et al. 2012). Therefore, in principle, some performance differences are expected.

In order to identify the GCs, the most interesting measure is the purity, i.e. the fraction of true GCs within the set of objects that are classified by the method as GCs (Brescia et al. 2012), although we are interested to find the best trade-off with completeness, as usual in any classification scheme but crucial in astrophysical problems (D’Isanto et al. 2016).

Concerning the *ugri* data set, both models show comparable performances in terms of trade-off between purity and completeness of the order of 80–95 per cent (see Table 1). However, in order to explore the classification differences between the two methods, we counted the number of times a model exceeds the other for more than 3 per cent in terms of statistical measurements. Concerning the *ugri* data set, MLPQNA overcomes the GNG results in the *GCSTAR* experiment (showing an average improvement of  $\sim 5$  per cent), while GNG performs better than MLPQNA in the *GCALL* experiment (where the GNG average increase is  $\sim 3$  per cent). Regarding the *3CLASS* experiments, both methods show very similar efficiency (in fact, the MLPQNA average increase is less than 0.4 per cent). Furthermore, GNG seems to identify GCs better than other classes (showing an average improvement of  $\sim 2$  per cent in the case of the *ugri* experiments), while MLPQNA shows a greater detecting capability for stars and galaxies (with respect to which the average gain related to *ugri* experiments is  $\sim 3$  per cent).

The scenario is different for the *gri* data sets, where MLPQNA outperforms GNG, by disentangling the source classes with very few losses (by reaching a purity and a completeness on GCs

**Table 1.** Classification results in terms of statistical estimators: average efficiency (*AE*), purity (*pur*), completeness (*compl*), and F1-score (*F1*) for both *ugri* and *gri* data set types and for both GNG and MLPQNA models. Top table reports the results for the *3CLASS* experiment, middle table for the 2-class experiment between GCs and *notGCs* (stars + galaxies), and bottom table shows the results concerning the 2-class experiment between GCs and stars. Regarding the *ugri* data set case, GNG and MLPQNA have similar performances, although MLPQNA shows an optimal trade-off between purity and completeness; while in the case of *gri* data set the GNG cannot reach MLPQNA performance, particularly in the 3-class and stars versus GCs cases. All these experiments are performed using the *BEST* PS. The values higher than 90 per cent are marked in bold.

<i>3CLASS</i>		<i>ugri</i>		<i>gri</i>	
Estimator (per cent)	GNG	MLPQNA	GNG	MLPQNA	
AE	86.5	88.2	79.4	88.8	
pur STAR	85.8	84.8	71.9	85.6	
compl STAR	80.3	85.6	66.9	82.0	
F1 STAR	83.0	85.2	69.3	83.8	
pur GCs	80.0	83.2	78.2	87.2	
compl GCs	<b>90.8</b>	83.2	79.6	89.4	
F1 GCs	85.1	83.2	78.9	88.3	
pur GAL	<b>92.5</b>	<b>95.4</b>	88.3	<b>94.5</b>	
compl GAL	<b>95.4</b>	<b>94.7</b>	<b>92.1</b>	<b>94.7</b>	
F1 GAL	<b>93.9</b>	<b>95.0</b>	<b>90.2</b>	<b>94.6</b>	
GCs versus ALL		<i>ugri</i>		<i>gri</i>	
Estimator (per cent)	GNG	MLPQNA	GNG	MLPQNA	
AE	88.7	87.1	84.0	88.4	
pur notGC	85.1	<b>91.2</b>	81.3	<b>90.1</b>	
compl notGC	88.0	89.6	85.5	89.4	
F1 notGC	86.5	<b>90.5</b>	83.4	89.8	
pur GCs	<b>91.3</b>	78.9	86.2	86.3	
compl GCs	89.1	81.8	82.1	87.1	
F1 GCs	<b>90.2</b>	80.3	84.2	86.7	
GCs versus STARS		<i>ugri</i>		<i>gri</i>	
Estimator (per cent)	GNG	MLPQNA	GNG	MLPQNA	
AE	86.8	<b>90.3</b>	78.2	87.9	
pur STAR	87.0	84.3	77.3	81.8	
compl STAR	83.2	80.7	84.9	83.2	
F1 STAR	85.1	82.5	81.1	82.5	
pur GCs	<b>91.6</b>	<b>92.6</b>	79.7	<b>91.2</b>	
compl GCs	80.3	<b>94.1</b>	70.3	<b>90.4</b>	
F1 GCs	85.6	<b>93.3</b>	75.0	<b>90.8</b>	

higher than 86 per cent), although without the *u* band. Only in the *2CLASS* experiment *GCs versus ALL* the GNG network achieves similar (although minor) performances. In this case, the AE difference between model classification capabilities is  $\sim 4$  per cent, whereas in *3CLASS* and *GCSTAR* experiments the differences are  $\sim 10$  per cent and  $\sim 8$  per cent, respectively.

By comparing the results between the *GCSTAR* and the other experiments, it appears that the exclusion of galaxies from the train set makes both models more efficient to identify the GCs, although both methods are able to identify galaxies with a good trade-off between purity and completeness ( $\gtrsim 92$  per cent in the *ugri* case,  $\gtrsim 88$  per cent for the *gri* experiment). This behaviour appears more pronounced for the MLPQNA.

The GNG performance gaps between *ugri* and *gri* data set can be visually deduced also from the ROC curves (Fig. A1) studying

**Table 2.** Intersection between predictions performed by GNG and MLPQNA for both *ugri* and *gri* data set types. Further, in order to disentangle the influence on performances due to the amount of samples in the train set, we evaluate the intersection between the predictions performed on the *gri* data set using only the *ugri* indices (i.e. *gri* features with *ugri* samples), named as *gri\**. Top table refers to the *3CLASS* experiment, middle table to the *GCALL* experiment, while bottom table to the *GCSTAR* experiment. Row *commons* reports the intersection between predictions regardless of whether they are correct or not. Row *corrected* specifies the common objects correctly classified. The values are expressed: (i) as percentage with respect to whole set (third and fifth columns) and (ii) as percentage with respect to the number of objects in the corresponding class (fourth and sixth columns).

		3CLASS					
		<i>ugri</i> (per cent)	<i>gri</i> (per cent)	<i>gri*</i> (per cent)			
Commons	STAR	85.3	78.8	80.8	67.3	78.9	72.1
	GCs		79.8		82.4		67.6
	GAL		95.2		92.1		92.0
Corrected	STAR	80.1	75.9	75.0	63.0	75.6	77.6
	GCs		71.1		74.4		58.0
	GAL		91.3		88.2		88.9
		GCALL					
		<i>ugri</i> (per cent)	<i>gri</i> (per cent)	<i>gri*</i> (per cent)			
Commons	notGCs	86.0	88.3	82.2	81.1	83.0	85.6
	GCs		82.6		83.4		78.7
Corrected	notGCs	80.6	84.6	77.0	77.1	77.0	81.9
	GCs		75.7		77.1		71.1
		GCSTAR					
		<i>ugri</i> (per cent)	<i>gri</i> (per cent)	<i>gri*</i> (per cent)			
Commons	STAR	90.3	78.6	80.0	70.4	80.6	82.8
	GCs		95.0		86.2		74.7
Corrected	STAR	84.7	74.3	73.1	66.0	72.3	77.8
	GCs		88.8		77.8		66.7

the trend of the area under the curve (AUC), which represents the probability that a classifier correctly predicts the membership of a sample, although the ‘positive’ probability thresholds are higher than the ‘negative’ ranks (Fawcett 2006). Concerning the GC classification, the AUCs gain up to 9.9 per cent by moving from *gri* to *ugri* data set; the best result is obtained by the *GCALL-ugri* experiments (93 per cent and 94 per cent, respectively, for GCs and *notGCs*). The performances drop down for the *GCSTAR-gri* experiment (AUC  $\sim$  80 per cent for both classes), where the photometric similarity between sources and the lack of information makes the GNG less performing.

In Table 2, we distinguish the commonalities between GNG and MLPQNA referred to their predictions and corrected ones. Such analysis was computed for all the six kinds of experiments and evaluated on a PS composed by the *gri* features, but restricting the data samples only to the objects available in the *ugri* case (named as *gri\**). Again, the underlying idea is to disentangle the contribution of features from the increasing of sample size in the train set. As expected, from the estimated performances (see Table 1), the largest sets of common predictions are related to the *ugri* experiments (80–90 per cent). As previously discussed, the inclusion of galaxies in the training set reduces the capabilities of both methods to detect GCs (a 5 per cent drop in term of commonalities). Concerning the *gri\** case, there is a negligible reduction in terms of global common classification fraction (i.e. regarding all the involved source types), but the fraction of identified (and correctly identified) GCs is

considerably decreased (up to 20 per cent). In order to understand the origin of the misclassified objects, we plot on colour–colour diagrams the spectroscopic sources together with the incorrect predictions with respect to all classes of objects, for both *ugri* and *gri* training sets, shown in Fig. 3. The incorrect sources could be due to the photometric similarity between the sources, for instance GCs and stars, accentuated when the *u* band is removed. However, given the high number of dimensions involved, the diagrams describe a small portion of the whole space of features (each one of them represents less than the 5 per cent of total information contribution, estimated as sum of the feature importances related to the involved colours). Thus, we used our model *ECODOPS*<sup>4</sup> (Efficient Coverage of Data On Parameter Space), a PYTHON-based system wrapping a high-dimensions data visualization technique: t-distributed Stochastic Neighbor Embedding (tSNE; van der Maaten & Hinton 2008; Van Der Maaten 2014). Our tSNE implementation is based on the object imported from the library *sklearn* (Pedregosa et al. 2011).

Such method, already applied in other astrophysical contexts (Nakoneczny et al. 2019), guarantees the preservation of the significant structures of the high-dimensional data visualized in a low-dimensional map. Therefore, it converts similarities between data points to joint probabilities and tries to minimize the Kullback–Leibler (Kullback & Leibler 1951) divergence between the embedding space and the high-dimensional data. In this way, the tSNE maps the multidimensional data to a lower dimensional space and attempts to find patterns in the data by identifying observed clusters based on similarity of data points with multiple features. However, after this process, the input features are no longer identifiable, and any inference based only on the output of the method cannot be done. Hence, it must be considered as mainly a data exploration and visualization technique. The resulting 2D representation is obtained as a bunch of data points scattered on a 2D space (Fig. 4), where the underlined concept is that two close data points in the 2D space have similar properties in the high-dimensional space.

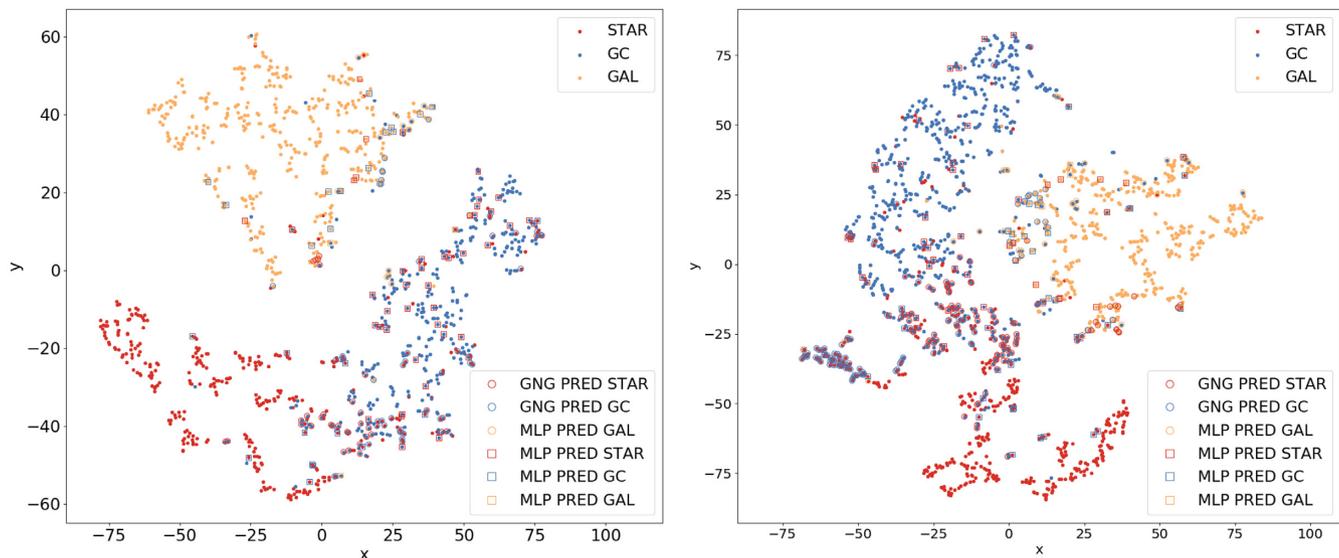
The embedding maps, illustrated in Fig. 4 for both *ugri* and *gri* sets, together with the misclassified objects, show the great separability between the class types when the information carried by the *u* band is added to the train set. Most of the incorrect predictions are located in the border regions, particularly in the *ugri* case (left-hand panel in Fig. 4), while other false predictions are surrounded by contaminants. The embedded space shows a correspondence to the colour–colour plane: the similarity between stars and GCs is still noticeable, although their separation is larger than in any other feature combinations, i.e. the whole ensemble of features computed by  $\Phi$ LAB maximizes the separation capabilities of both methods.

### 4.3 Photometric search for new GCs in Fornax core

It is now crucial to verify the capability of GNG to identify GCs by testing it against a set of unlabelled sources (hereafter, *run* process). This should allow not only to analyse the performances of the classifiers, but also to provide an additional set of GCs, suitable for advances in the astrophysical studies of the GC population and their connection with host galaxies.

As preliminary step, the fainter sources are excluded from the data set, by applying the cuts summarized in Table 3, due to their low S/N ratio and in order to cut unclassified data at the same

<sup>4</sup><http://dame.dsf.unina.it/ecodops.html>



**Figure 4.** Bi-dimensional projection performed by *ECODOPS* of the *ugri* (left-hand panel) and *gri* (right-hand panel) train set together with the misclassified sources predicted by GNG (open circles) and MLPQNA (open squares). In both figures, GCs (spectroscopic and predicted) are in blue, stars (spectroscopic and predicted) are in red, and galaxies (spectroscopic and predicted) are in orange.

**Table 3.** Magnitude and magnitude error cuts adopted for the run data set, for the *u*, *g*, *r*, and *i* bands, deduced from error trends.

Magnitudes	<i>u</i> -band	<i>g</i> -band	<i>r</i> -band	<i>i</i> -band
<i>AUTO</i>		23.7	23.0	23.0
<i>APER4</i>		25.2	25.0	24.6
<i>APER6</i>		24.2	24.1	23.6
<i>APER8</i>		24.3	23.6	23.0
Errors	<i>u</i> -band	<i>g</i> -band	<i>r</i> -band	<i>i</i> -band
<i>AUTO</i>	0.18	0.040	0.040	0.050
<i>APER4</i>	0.04	0.050	0.050	0.070
<i>APER6</i>	0.05	0.035	0.034	0.055
<i>APER8</i>	0.18	0.033	0.030	0.050

limit of the KB. The *u*-band magnitudes were excluded from the cut, since the training and the run magnitude distributions share the same range of values. Moreover, cuts on magnitude errors have been applied on all the available bands, in order to limit the presence of noisy sources (Table 3). Furthermore, samples affected by missing data were excluded from the catalogue. At the end of this selection process, two data sets have been produced, one including the *u* band, the other excluding it, in a similar way to what was done for the KB data. The *ugri* data set consists of 5562 sources ( $\sim 45$  per cent of the available run set), while the *gri* data set counts 6884 sources ( $\sim 17$  per cent of the available run set).

We have performed the *run* process with the GNG whose learning had involved all the three source types, i.e. galaxies, stars, and GCs. Thus, the GNG models trained with the *GCSTAR* have been excluded from the *run* process. Indeed the *run* set is composed by all sources detectable from the instrument, so, when a galaxy is presented to this network, the model tries to assign a label to the source, i.e. *star* or *GC*, making a mistake in both cases. The purpose of the *GCSTAR* experiments is to test the effectiveness of the network to photometrically disentangle GCs from stars, which is the most complex among the proposed problems, due to the morphological and photometric similarity of both source types.

**Table 4.** Common predictions among the GNG trained networks performed on the unlabelled sources for the four data set experiments involving the three class types: upper rows refer to *3CLASS* experiments, while bottom rows refers to *GCALL* experiments.

<i>3CLASS</i>	COMMON	Per cent	GCs	Stars	Galaxies
<i>ugri</i>	5115	92.0	522	2022	2571
<i>gri</i>	5861	85.1	425	2601	2835
<i>GCALL</i>	COMMON	Per cent	GCs	notGC	
<i>ugri</i>	5228	94.0	472	4756	
<i>gri</i>	6437	93.5	790	5647	

Since a *leave-k-out* approach has been adopted, we used the five available trained networks to analyse the GNG performance fluctuations. Table 4 shows the results of the intersection between the different results produced by the GNG in terms of common predictions among stars, GCs, and galaxies sources: as expected from the blind test performance (Table 1), the common percentages reveal a gap between the *ugri* and *gri* *3CLASS* experiments; nevertheless the other three run experiments reach more than 90 per cent of common predictions, finding about 500 GCs candidates.

After having verified the robustness of the method with respect to the data set variations and that the results seem to reflect the performances achieved on the blind test set, we trained the GNG on the whole KB. In order to quantify the overall performances of the network, we used samples of *bona fide Hubble space Telescope (HST)* GCs in the central region of *NGC 1399* (Brescia et al. 2012; Cavuoti et al. 2013; Puzia et al. 2014). After a cross-match between VST and *HST* catalogues, we found 100 *HST* sources (GC and *notGC*) within our run data set. The resulting classification statistics are shown in Table 5. Despite the reduced number of samples, the measures reflect what was obtained on the blind tests: an increase of the classification accuracy for the *ugri*-band data set with respect to the *gri* case, and for the *GCALL* experiments with respect to the *3CLASS* case.

We want to emphasize the result obtained with the experiment *ugri GCALL*, which outperforms the others, reaching an excellent

**Table 5.** Classification results in terms of statistical estimators using *HST* samples as *bona fide* for both *ugri* and *gri* data sets. The estimator nomenclature is the same as that adopted in Table 1. The columns *3CLASS* shows the results concerning the 3-class experiments, while the columns *GCALL* shows the results concerning the 2-class. Despite the limited amount of labelled sources within the *HST* sample, the statistical estimators reflect the performance obtained with the blind test (Table 1).

Estimator (per cent)	<i>ugri</i>		<i>gri</i>	
	<i>3CLASS</i>	<i>GCALL</i>	<i>3CLASS</i>	<i>GCALL</i>
AE	90.1	96.7	80.4	85.5
pur <i>notGCs</i>	92.7	97.7	86.2	89.1
compl <i>notGCs</i>	86.4	95.6	64.1	81.7
F1 <i>notGCs</i>	89.4	96.7	75.1	85.4
pur GCs	88.0	95.6	77.8	82.3
compl GCs	93.6	97.8	92.5	89.1
F1 GCs	90.8	95.6	85.1	85.7

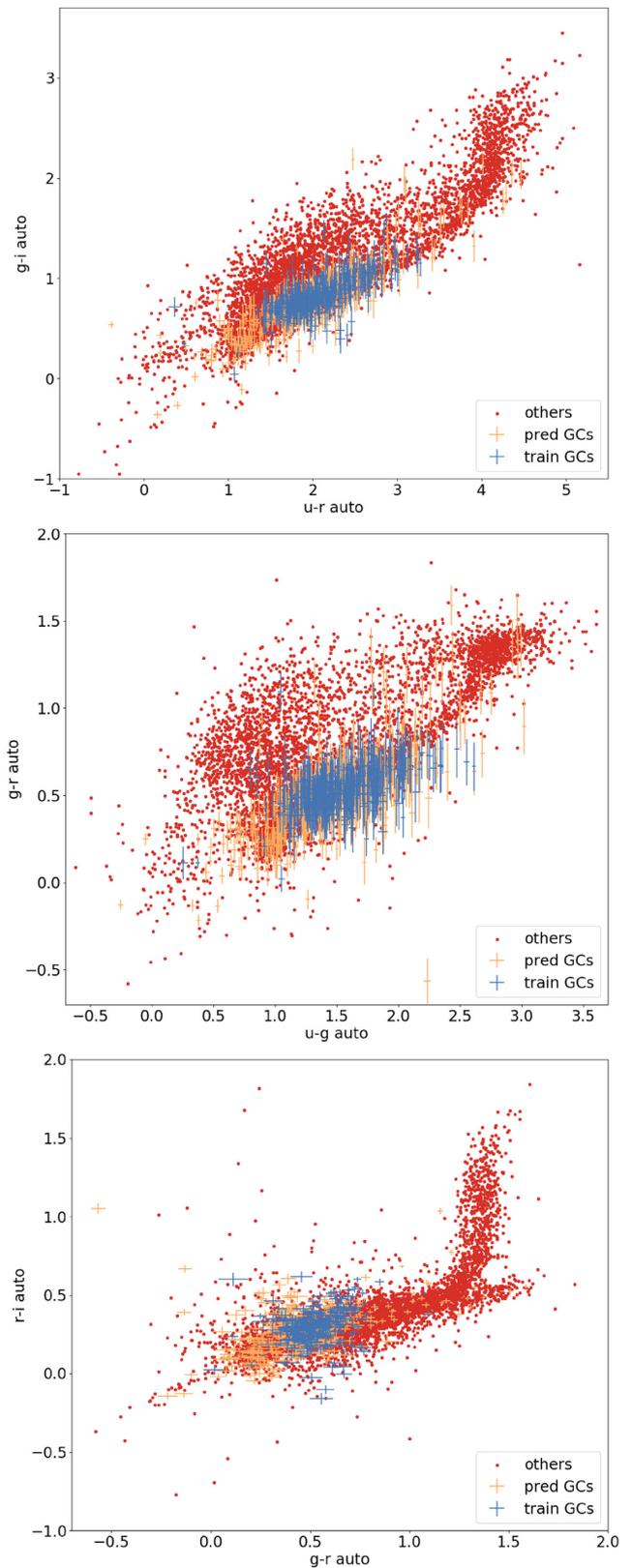
purity-completeness trade-off and a very low GC–*notGC* contamination ( $\lesssim 4$  per cent). Thus, given these measures and the performances achieved on the blind test set, we used the GCs identified by the GNG trained with *ugri GCALL* data set. Furthermore, since we were also interested in other source types, we defined as stars (or galaxies) the sources classified as *notGCs* trained with the *ugri GCALL* experiments, which have been predicted as stars (or galaxies) by GNG trained with the *ugri 3CLASS* experiment, i.e. the common prediction among the *ugri* experiments.

The resulting colour–colour diagrams for the predicted GCs are illustrated in Fig. 5 together with the other sources (i.e. galaxies and stars). These panels clearly show a large overlap between predicted and training GCs probing the network capability to extract the GC population, photometrically indistinguishable from the background and foreground sources. The results confirm the capability of our method, able to identify the sources without any pre-selection and in spite of the limited number of labelled sources. The only requirement is the correspondence, in term of photometric coverage, between the training and the run data sets.

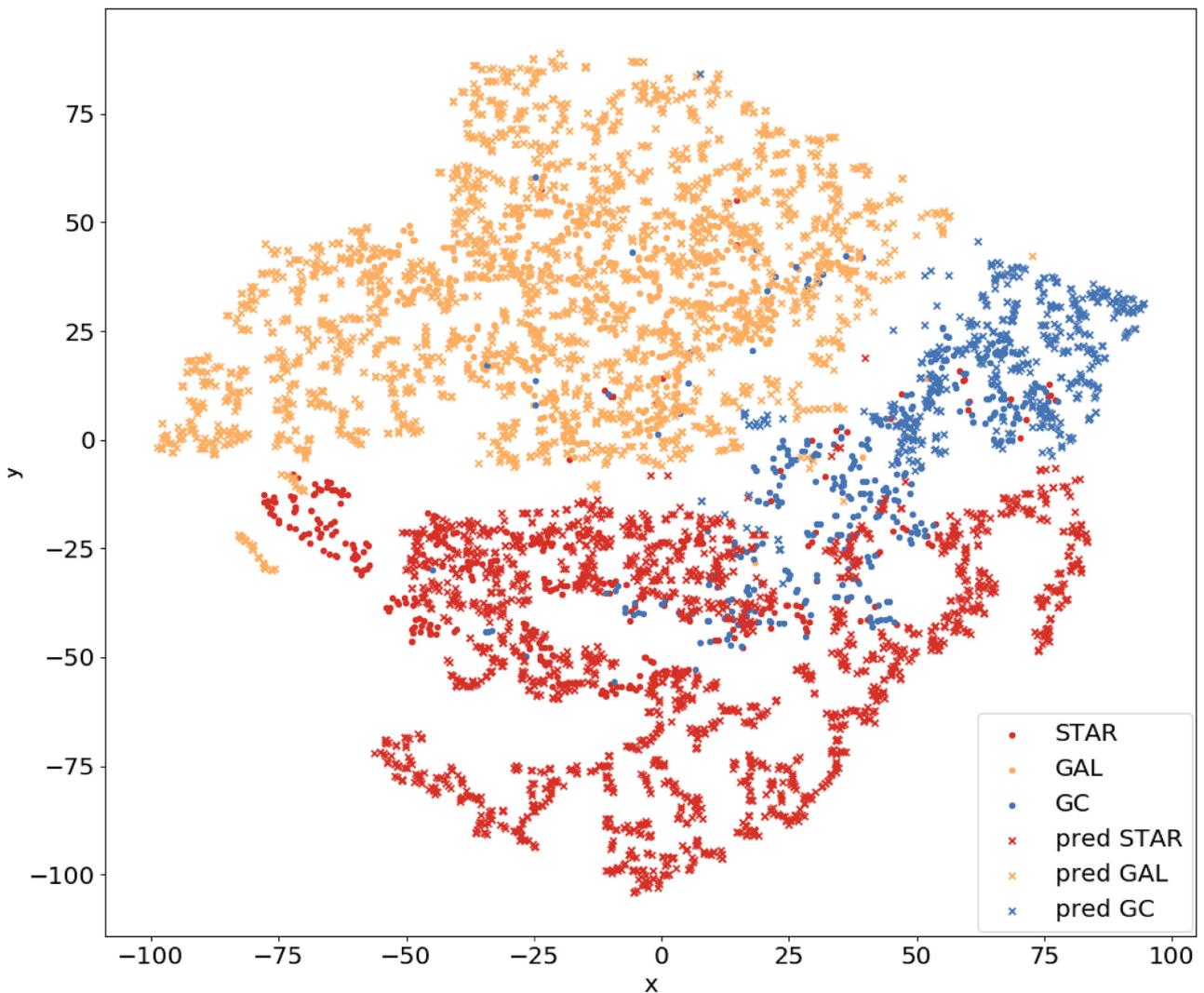
Concerning the residual misclassified objects, some of these could be false positives (FPs). This misclassification could be due to the exiguous number of training sources or a non-uniform sampling of the PS. In the left-hand panel of Fig. 4, we have shown a visualization of the *ugri* train set into a bi-dimensional space through the *ECODOPS* tool. Although this is a projection, it is evident that the space is not uniformly sampled and characterized by the presence of several contaminants. Such two factors, together with the exiguous number of training sources, could cause the presence of the outliers. Thus, in order to visualize the result of the run process, we estimate the bi-dimensional projection of the run set, analogous to what was already done in Section 4.2. In Fig. 6, we show this same projection by overlapping the training set objects. Most of the predicted sources seem to populate well-defined regions, predominantly occupied by spectroscopic objects, although stars and GCs show a large overlap, as expected. This aspect, together with the already discussed problems of the misclassification at the border of class regions, may cause the presence of redder and bluer outliers.

Given such premises, by considering also the low fraction of spectroscopic sources available, a larger fraction of FPs could be expected. However, the resulting exiguous number of contaminants is a consequence of the approached PS optimization process.

Clearly these outliers must be excluded from the set of the identified GCs. A simple approach could be excluding sources from the colour–colour diagrams. However given the arbitrariness



**Figure 5.** Colour–colour diagrams for the run predicted sources (orange) compared with the train sources (blue), overlapped to the colour–colour distribution related to the other sources (i.e. galaxies and stars, red dots in figures). From the top panel to the bottom the figures refers to the colours, respectively,  $g-i$  versus  $u-r$ ,  $g-r$  versus  $u-g$ , and  $r-i$  versus  $g-r$ .



**Figure 6.** Bi-dimensional projection performed by tSNE of the *ugri* set both for spectroscopic sources (blue, red, and orange dots, respectively, for GCs, stars, and galaxies) and for predicted candidate objects (blue, red, and orange crosses, respectively, for GCs, stars, and galaxies).

of such procedure, we neglected this solution. In order to exclude such candidate FPs, we took advantage from the existence of the two GC populations (namely red and blue, Kundu & Whitmore 1998), to fit a bimodal bivariate Gaussian distribution deduced from the spectroscopic GCs. We used a Gaussian Mixture Model (GMM, Muratov & Gnedin 2010) implemented through the library *sklearn* (Pedregosa et al. 2011), which is a generalization of the *K*-Mixture Model (Ashman, Bird & Zepf 1994). The method maximizes the likelihood of the data set using the EM algorithm, which allows to derive explicit equations for the maximum-likelihood estimate of the parameters. The projection of this surface on the colour–colour plane is illustrated in the top panel of Fig. 7. The red and blue ellipses symbolize the contour levels matching, respectively,  $1\sigma$ ,  $2\sigma$ , and  $3\sigma$  of the bivariate bimodal Gaussian distribution of the underlying spectroscopic GC population. The black line crossing the ellipses is the projection on the colour–colour plane of the intersection between the two bivariate Gaussian surfaces. Training (i.e. spectroscopic) and predicted sources above such line and within the  $3\sigma$  levels are considered components of the red population, while those below and within the  $3\sigma$  levels are assumed to be members of the blue population. Finally, we assume as FPs the predicted sources outside the union of the  $3\sigma$  ellipses (110 objects,  $\sim 23.3$  per cent).

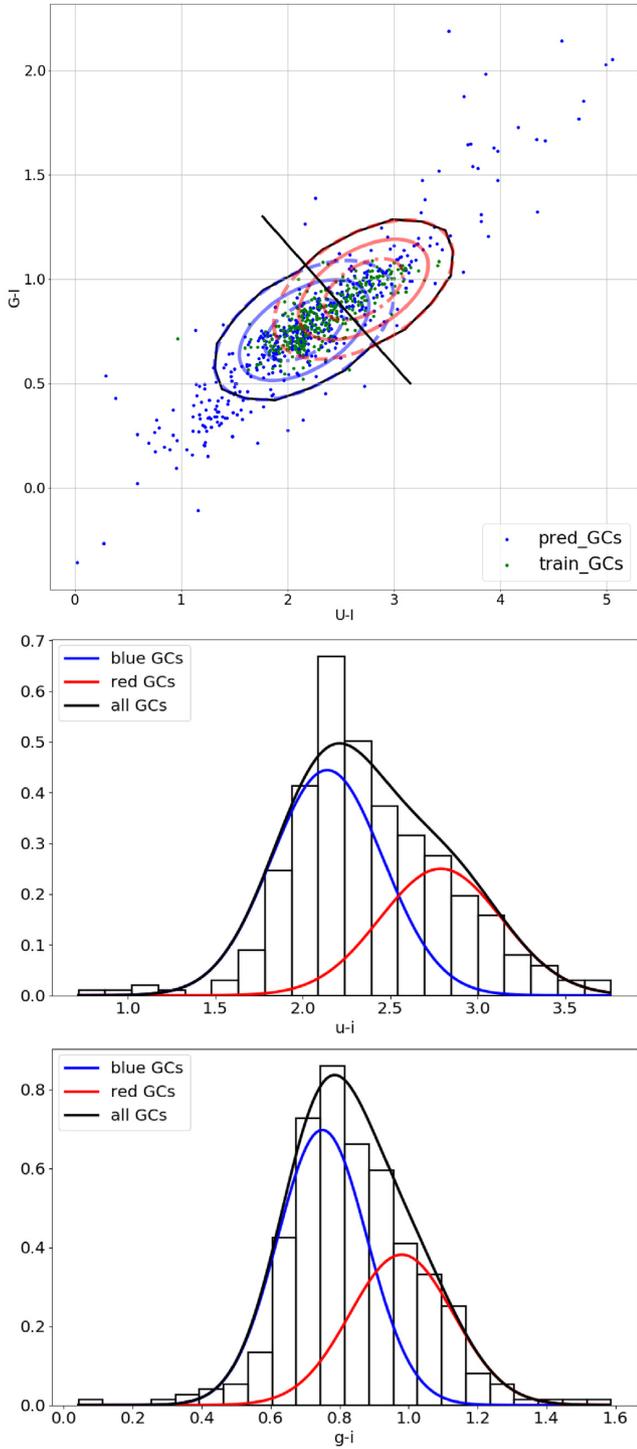
Middle and bottom panels of Fig. 7 show the evident bimodal colour distributions of the GC population. Once the FPs have been excluded from the GC set, the intracluster error (Floudas & Pardalos 2006; Murtagh & Legendre 2014), defined as the measure of the overlap between training and predicted GCs, decreases by  $\sim 5$  per cent while variance drops down by about 80 per cent.

Panels in Fig. 8 illustrate the colour–colour diagrams for the selected GCs, galaxies, and stars. Concerning the branch of the stars, some bluer sources could be false positives, since their distribution appears to be the extension of the galaxies trend. However, the stellar branch goes through the diagrams following the expected shapes and, above all, the GCs occupy the restricted region well known in literature (e.g. Cantiello et al. 2018a), in which GCs are particularly difficult to be separated from stars or galaxies through colours.

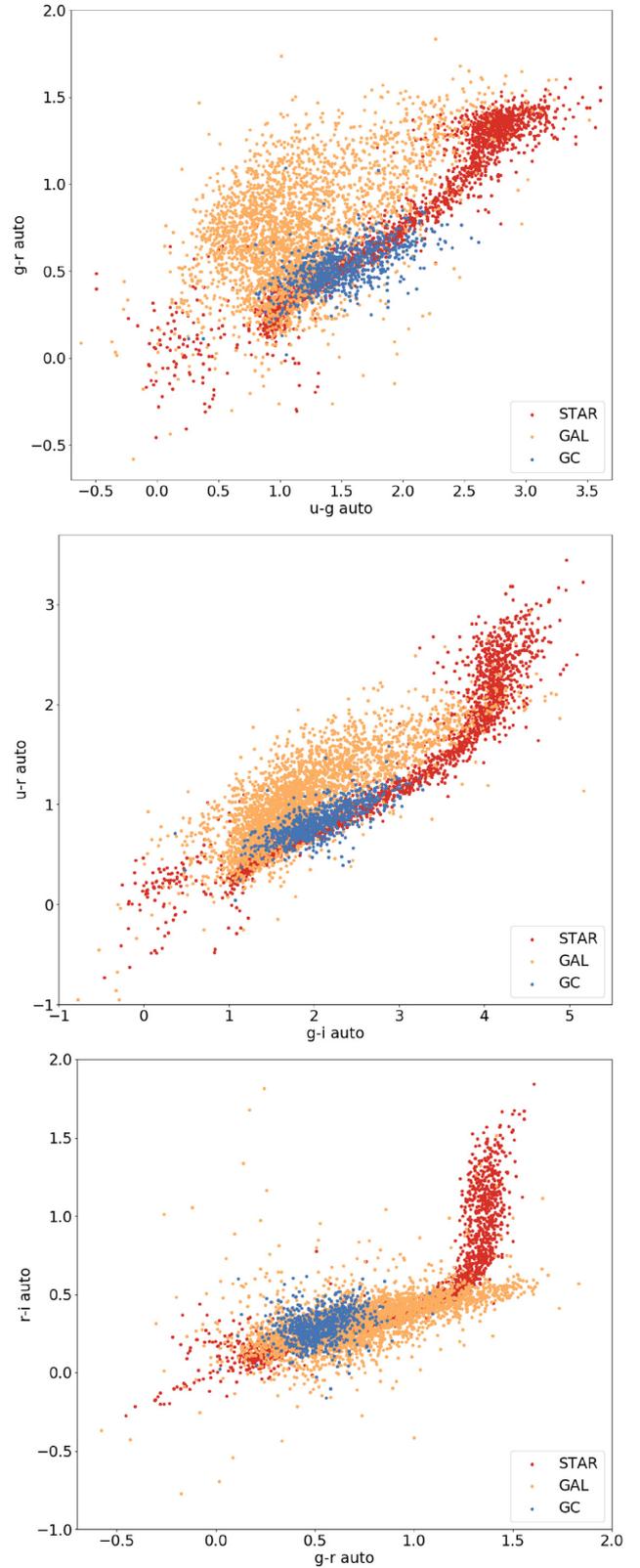
## 5 COMPARISON WITH EXTERNAL DATA

In order to validate the GCs identification through the GNG-GMM approach, we compare our selection with other similar works:

- (i) ML experiments performed by Brescia et al. (2012) and more recently by Angora et al. (2017) on single-band *HST* data of *NGC*



**Figure 7.** Top panel: Projection of the bivariate bimodal Gaussian distributions on the colour–colour plane fitting the training GCs (in green). The predicted GCs are scattered in blue. The ellipses are the contours levels related to  $1\sigma$ ,  $2\sigma$ , and  $3\sigma$ , respectively, for the red and blue population. All sources within the union of the  $3\sigma$  ellipses are assumed to be *true* GCs. The black line is the projected intersection between the bivariate Gaussians. Middle and bottom panels: Colour distribution ( $u-i$  and  $g-i$ ) for the bimodal GC population.



**Figure 8.** Colour diagrams for stars, galaxies, and GCs downstream the GMM false positive exclusion process. From the top panel to the bottom one the figures show:  $g-r$  versus  $u-g$ ,  $r-i$  versus  $u-g$ ,  $r-i$  versus  $g-r$ . In these figures, the GCs are blue, the stars are red, and the galaxies are orange. The diagrams recall the results obtained in others works (e.g. Cantiello et al. 2018a).

**Table 6.** Comparison between classification performed with *HST* and VST data. The results related to the *HST* data obtained by the MLPQNA and GNG are derived, respectively, from Brescia et al. (2012) and Angora et al. (2017). The estimators are referred to the GCS. In order to measure the amount of correctly classified GCs, the contamination (dual estimator of the purity) has been used.

	Estimator (per cent)	MLPQNA	GNG
<i>HST</i>	AE	98.3	86.8
	Completeness	97.8	83.8
	Contamination	1.6	15.9
VST	AE	87.1	88.7
	Completeness	81.8	89.1
	Contamination	21.1	8.7

1399. This comparison, presented in Section 5.1, analyses the performances achieved by the same methods (GNG and MLPQNA) varying the instruments (VST versus *HST*);

(ii) Other experiments carried out by D’Abrusco et al. (2016) and Cantiello et al. (2018b), which exploit techniques different from ML methods on the same VST data set. In this case, we compare performances obtained by different approaches (ML versus not-ML) using the same instruments (discussed in Section 5.2).

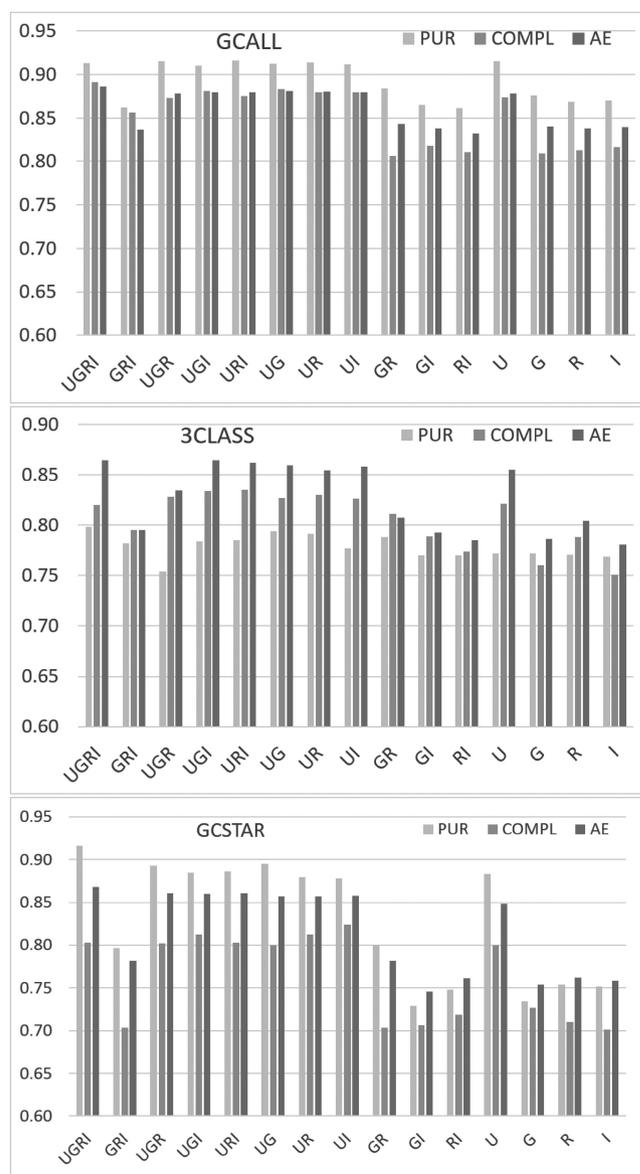
### 5.1 Comparison with *HST* data

The catalogue used by Brescia et al. (2012) and Angora et al. (2017) was extracted from single-band *HST* images of *NGC 1399*, reaching  $7\sigma$  at  $m_V = 27.5$ , that is  $\sim 3.5$  magnitudes fainter than the GC luminosity function turnover point, thus it allows the sampling of nearly the entire GC population (Puzia et al. 2014). The PS is composed by seven photometric features, respectively, four magnitudes (isophotal and three different apertures), *FWHM*, *central surface brightness*, *Kron radius* and four structural parameters, respectively, *King’s tidal*, *effective radius*, *core radius*, and *ellipticity*.

Table 6 reports a comparison between the best results obtained by the MLPQNA and the GNG networks on the *HST* and VST data (the latter extracted from Table 1). Instead of the purity, the contamination (i.e. the complementary of purity,  $1 - \text{purity}$ ) was used to evaluate the capability of the ML models to correctly classify the GCs. MLPQNA achieves a remarkable result on the *HST* data, with a contamination of  $\sim 1.8$  per cent (thus corresponding to a purity of  $\sim 98.2$  per cent). The GNG performances appear to be similar by increasing the accuracy in the case of VST data. The statistics suggest the capability of ML models to disentangle the GCs from the background and foreground sources also with high-quality single-band photometry. In order to investigate such result, the *k*-fold-based training/test procedure, described in Section 4, has been reproduced for VST data using all possible filter combinations and varying the involved number of bands for all the classification experiments.

Panels in Fig. 9 show the performances of the GNG model as a function of the used band filters, probing, as expected, the effectiveness of the complete spectrum data set in order to identify GCs with ground-based imaging.

The *ugri* case has the best trade-off between purity and completeness (related to the GCs) in the three classification experiments, i.e. the model is able to correctly identify the GCs with an acceptable level of precision and sensitivity. Since the purity and the completeness are referred to GCs, the average efficiencies have been displayed in Fig. 9, in order to include classification information about the *notGCs*. Concerning the incomplete spectrum



**Figure 9.** Comparing among GNG network blind test performances as function of the involved filter numbers in terms of GCs purity (*pur*, light grey), completeness (*compl*, grey), and average efficiency (*AE*, dark grey) for the experiments *GCALL* (top panel), *3CLASS* (middle panel), and *GCSTAR* (bottom panel). Experiments involving the *U* band improve significantly their purity (more than 5 per cent in the *GCALL* experiment), their completeness (more than 7 per cent in the *GCALL* experiment), and their average efficiency. This latter has been included to add information about the *notGCs* classification.

data sets, the experiment involving the *U* band improves the purity by  $\sim 0.6$  per cent to  $\sim 6.3$  per cent, the completeness by  $\sim 3.1$  per cent to  $\sim 7.7$  per cent, and the average efficiency by  $\sim 3.8$  per cent to  $\sim 9.4$  per cent. Therefore, there is a significant performance gap between experiments with or without the *U* band.

Comparing the accuracy reached by the GNG network on the *HST* and the single-band VST data, the experiment based on the *u* band is the only one outcoming comparable results. Only using the complete spectrum the GNG achieves better performance than single-band *HST* GNG experiments, improving purity, completeness and

efficiency (by  $\sim 8.3$  per cent,  $\sim 5.3$  per cent, and  $\sim 1.8$  per cent, respectively, for the *GCALL* experiment).

Brescia et al. (2012) probed the capability of several ML methods to disentangle GCs from background and foreground sources using single-band, high-quality, deep photometry. To achieve similar results on ground-based VST data, it is necessary to use at least two filters, where one of them has to be the *u* band.

Regarding the feature FS procedure, it is possible to compare results obtained by Angora et al. (2017) with *HST* data, which added an FS procedure provided by the RF model, finding a set of relevant features composed by only photometric quantities, from which the *Kron radius* and the *ellipticity* were also rejected. Despite some differences between the data sets, for instance, the magnitude coverage, depth, and number of bands, the FS results are similar to those performed through  $\Phi$ Lab. In fact, in both cases the *Kron radius* is rejected, showing a negligible informative contribution ( $< 1$  per cent). Concerning the *ellipticity*, defined as  $1 - B\_WORLD/A\_WORLD$ , it is related to another  $\Phi$ LAB rejected feature, the *ELONGATION*, connected to the *ELLIPTICITY* through  $ELLIPTICITY = 1 - (1/ELONGATION)$  (Bertin & Arnouts 1996). Therefore, its rejection is motivated by the informative contribution already carried by the *ELONGATION* feature.

## 5.2 Comparison with other techniques

With the term *other* we refer to those methodologies that do not exploit ML, and use instead several combinations of cuts in a more or less complex PS, to separate GCs from background/foreground sources. We have compared our results with two works, respectively, D’Abrusco et al. (2016) and Cantiello et al. (2018b), which analyse the Fornax region with the same VST images.

D’Abrusco et al. (2016) applied a PCA (Bishop 2006) on the natural colours, identifying a *locus* in the PC space dominated by the presence of GCs, excluding brighter and fainter sources (i.e. cuts on *G* band) and using the *SEXTRACTOR CLASS\_STAR* parameter.

Cantiello et al. (2018b) introduced a *morpho-photometric* approach: in order to analyse the properties of the GC sample (bimodality, density maps, radial profiles), they use a statistical background decontamination method. Here, for simplicity, we compare our results to the catalogue provided by Cantiello et al. (2018b), being aware that this is oversimplified. Cantiello et al. (2018b) used a spectroscopic set of sources in order to find the PS occupied by GCs and applied a set of cuts on the features:  $\Delta X$ , *CLASS\_STAR<sub>X</sub>*, *FWHM<sub>X</sub>*, *FLUX\_RADIUS<sub>X</sub>*, *KRON\_RADIUS<sub>X</sub>*, *PETRO\_RADIUS<sub>i</sub>*, *ELONGATION<sub>X</sub>*, *u-i*, *g-i*,  $\Delta(u-i)$ ,  $\Delta(g-i)$ ,  $\Delta m_X$ ,  $m_X$ . *X* labels the *g* and *i* bands, while  $\Delta$  indicates the difference between two apertures (with respect to magnitudes and colours), respectively, 6 and 12. In addition, they used a selection on the colour–colour plane:  $|(g-i) - [0.362(u-i) - 0.0205]| \leq 0.2$ . The high number of cuts derives from the need to reduce contamination introduced by peculiar sources. In order to study the colour bimodality, they selected GCs inside several annular regions, concentric on *NGC 1399*. For each GC set within the annular region they applied a GMM in order to fit a bimodal univariate Gaussian distribution using the *u-i* and *g-i* colour. Moreover, they compared the bimodal and the unimodal distributions, statistically validating the colour bimodality.

Regarding the split between blue and red GCs, the blue GCs of D’Abrusco et al. (2016) are the sources whose *g-i* colours are less than 0.85, while the blue GCs of Cantiello et al. (2018b) are those sources with *u-i*  $< 2.5$ . These thresholds are stated by authors in their respective works.

Table 7 shows the bimodal Gaussian best fit together with the parameters estimated by D’Abrusco et al. (2016), which applied a GMM method only for the colour *g-i* of a larger GC set (their VST catalogue covers  $\sim 8.4$  deg<sup>2</sup>), and by Cantiello et al. (2018b), which refer to an annular region whose radii are 2.5 and 5 arcmin, respectively.

Using the criterion of acceptability (Taylor 1996), we have estimated the number of standard deviations for which our measure ( $\mu_1$ ) differs from that of D’Abrusco et al. (2016) and (Cantiello et al. 2018b,  $\mu_2$ ), i.e.  $t = |\mu_1 - \mu_2|/\sigma$ . The resulting values are reported in the last row of Table 7. The discrepancies with the peak of the blue sub-populations are less than  $1.6\sigma$ , i.e. the measures are comparable. Concerning the red sub-populations, only the peaks related to *u-i* are compatible below  $1\sigma$ ; the *g-i* peaks differ from each other by at least  $3\sigma$ . However, we point out that in absolute terms the observed differences are small ( $< 0.1$  mag) and they could be easily explained by noticing that the different studies sample different galactocentric distances; in particular the value of Cantiello et al. (2018b) refers to an annulus within 5 arcmin from *NGC 1399* where the red GC component usually peaks at redder colours, and they did not use the *r* band, so the samples are inherently different in terms of possible contamination.

Fig. 10 shows the *U-I* versus *G-I* diagram related to the three sets of candidate GCs, and shows a large overlap between the sets. In Table 8, the common sources between our GC catalogue and the candidate GCs provided by D’Abrusco et al. (2016) and Cantiello et al. (2018b) are reported. There are, respectively, 91 per cent and 80 per cent GCs in common, of which 72 per cent and 80 per cent are blue, and 83 per cent and 66 per cent are red. This result confirms the capability of our method to predict the GC class type. Furthermore, among the excluded FPs, only 2 and 29 sources are candidate GCs, strengthening the FPs selection robustness, based on a GMM approach.

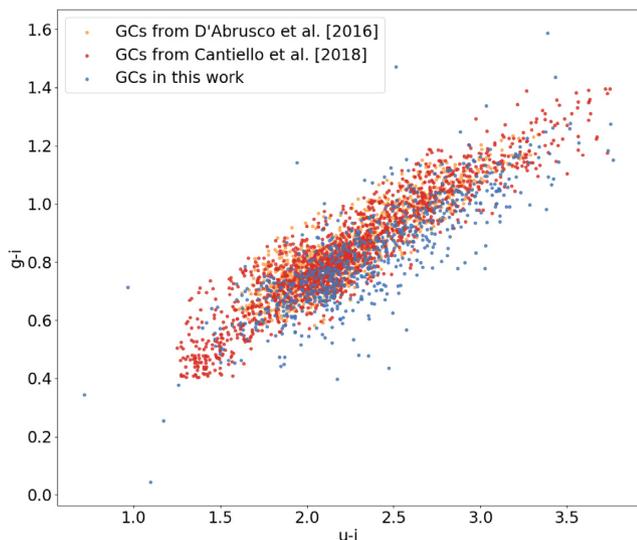
The resulting common sources reflect the difference between the selection approaches. All the cited works, included this one, use a GMM best fit to model a bimodal Gaussian underlying the GC populations, but through a different colour fitting: D’Abrusco et al. (2016) performed the fit with the *G-I* colour; Cantiello et al. (2018b) uses both *U-I* and *G-I* colours separately, in this work we used both colours by modelling a bivariate bimodal Gaussian distribution and producing a less sharp cut on the colour distributions.

Furthermore, we intersected the GCs provided by D’Abrusco et al. (2016) and Cantiello et al. (2018b) with both our training and predicted *notGCs*, i.e. stars and galaxies present in our training set together with those predicted by our method. From the intersection with the training stars, resulted 86 and 126 sources, respectively, for D’Abrusco et al. (2016) and Cantiello et al. (2018b), while only 2 and 1 sources resulted from the intersection with the training galaxies. Since the training *notGCs* are spectroscopically confirmed, these sources represent a set of candidate FPs for D’Abrusco et al. (2016) and Cantiello et al. (2018b).

In order to explore the differences between the GC populations and sub-populations, we estimated the cumulative distribution functions (CDFs) related to the *u-i* and *g-i* colours, illustrated in Fig. 11 for the whole GC population and for both the red and blue GC sub-populations, comparing our selected GCs with the selection performed by Cantiello et al. and D’Abrusco et al. Furthermore, we applied a Kolmogorov–Smirnov test (KS test; Peacock 1983; Fasano & Franceschini 1987) to estimate whether two samples have been extracted from the same distribution (here after null hypothesis). Concerning the whole GC systems, the largest

**Table 7.** Bivariatebimodal Gaussian parameters fitted through the GMM method compared with parameters fitted by D’Abrusco et al. (2016) and by Cantiello et al. (2018b). Values in the last row of the table refer to the discrepancies of distribution peaks that are consistent within  $(0.8, 3.1)\sigma$ . These discrepancies have been estimated as  $t = |\mu_1 - \mu_2|/\sigma$  (Taylor 1996).

	GMM				D’Abrusco et al.		Cantiello et al.			
	Blue		Red		Blue	Red	Blue		Red	
	$u-i$	$g-i$	$u-i$	$g-i$	$g-i$		$u-i$	$g-i$	$u-i$	$g-i$
$p$					0.63	0.37	0.52	0.53	0.48	0.47
$N$					1853	1095	78	79	71	70
$\mu$	2.14	0.75	2.79	0.98	0.74	0.95	2.08	0.78	2.74	1.06
$\sigma$	0.32	0.13	0.35	0.15	0.08	0.12	0.34	0.11	0.55	0.17
	$\sigma$ ’s discrepancy between peaks				1.1	3.0	1.6	1.1	0.8	3.1



**Figure 10.**  $U-I$  versus  $G-I$  diagrams related to our predicted GC (blue), and those predicted by D’Abrusco et al. (2016), orange and Cantiello et al. (2018b), red.

difference is found comparing the  $u-i$  distribution with Cantiello et al. (bottom left panel in Fig. 11), for which the  $p$ -value is  $<10^{-7}$ ; this could imply the rejection of the null hypothesis, nevertheless the same GCs distributions in  $g-i$  colours (top left panel in Fig. 11) show a very similar CDFs with a  $p$ -value  $\sim 4$  per cent, i.e. the null hypothesis cannot be rejected. This discrepancy could be due to the different magnitude spanning range of colours, for instance 1.5 mag for  $g-i$  and 3.5 mag for  $u-i$ , which could cause a higher relevance of outliers in the  $u-i$  case. Regarding the sub-populations (centre and right-hand panels in Fig. 11), the distributions are affected by the difference between the GC sets (due to the different GC identification techniques) and by the non-uniform blue–red selection criteria. Nevertheless the KS test between our and D’Abrusco et al. red GCs returns a  $p$ -value of 11.7 per cent, so we cannot reject the null hypothesis.

## 6 ANALYSIS OF DENSITY MAPS OF GC SPATIAL DISTRIBUTION

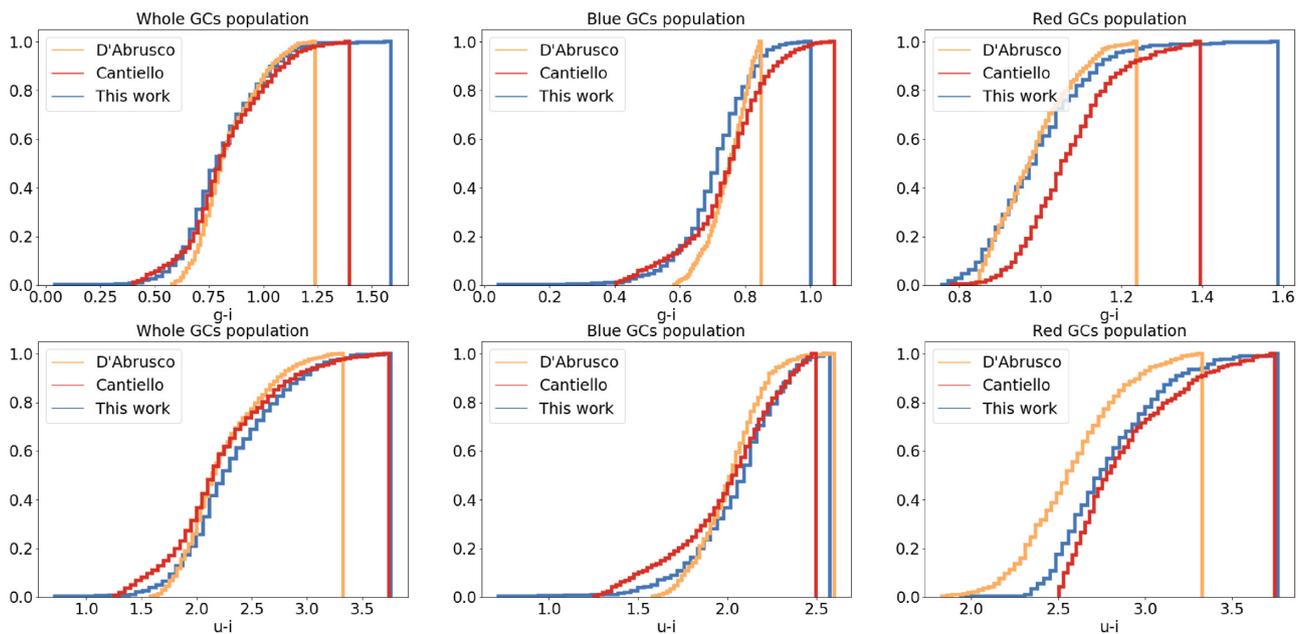
Finally, as further validation method, we present the density maps of the spatial distribution from the GC sky locations, estimated both for the whole population and for the blue and red sub-populations separately. The extracted density maps can be directly compared with those of D’Abrusco et al. (2016) and Cantiello et al. (2018b).

The density maps are related to the core of Fornax cluster, with  $RA \in (54.0, 55.0)$  and  $Dec. \in (-35.75, -35.13)$ , and have been estimated from the GCs sky coordinates, applying a  $K$ -nearest neighbour (KNN) method (Duda, Hart & Stork 2000) on a regular squared grid covering the sky region, following the same process presented in D’Abrusco Fabbiano & Zezas (2015) and D’Abrusco et al. (2016). Each knot in the grid has a density defined as  $d = K/(\pi \cdot r_k^2)$ , i.e. the ratio between the  $K$  neighbour GCs used to estimate the density, and the (projected) area of the circle whose radius is equal to the distance of the KNN. The value of  $K$  shapes the densities map: small  $K$ -values imply maps with compact density structures, while high  $K$ -values lead to large structures losing spatial information (D’Abrusco et al. 2015, 2016). In the following,  $K$  is taken equals to 9, adopting the same strategy proposed by D’Abrusco et al. (2016) which focused on the study of large spatial scale GC distribution. The GCs used in this process (719) are those used as training set for our GNG model and derived from the *run* executions after the GMM exclusion process. It is worth to say that the density is underestimated at the edges of the selected region, due to the lack of sources beyond those edges. Panels in Fig. 12 show the extracted density maps for the whole GC population (top panel), and for the blue and red sub-populations (middle and bottom panel, respectively). In these figures, the grey areas represent the region in which the density is underestimated.

Looking at the top panel in Fig. 12, the irregular shape of the region designed as *A* clearly shows a structure stretched in the W–E direction, due to the gravitational interaction between the giant elliptical galaxy *NGC 1399* (around which the density is maximum) and the nearby galaxies (*NGC 1396*, *NGC 1404*, *NGC 1387*, *NGC 1381*). This region contains 85 per cent of the involved GCs in the density map estimation. Within such region it is possible to distinguish an overdensity associated with the *NGC 1399–NGC 1396–NGC 1404* complex region (B, 46 per cent), where a bridge is connecting *NGC 1399* and *NGC 1404* in the SE–NW direction (discovered by Bassino et al. 2006 and emphasized by D’Abrusco et al. 2016). On a larger scale this complex region stretches to the west, combining densities related to *NGC 1387* and *NGC 1381*. At the north there is an isolated density region centred on *NGC 1380B* (C, 2.5 per cent). Iodice et al. (2017) have detected a previously unknown region of intracluster light (ICL). This overdensity of ICL is located in between the three bright galaxies in the core, *NGC 1387*, *NGC 1379*, and *NGC 1381*. They also show that the ICL is the counterpart in the diffuse light of the known overdensity in the population of blue GCs. A detail of the connection between *NGC 1387* and the complex region *NGC 1399–NGC 1396–NGC 1404* is illustrated in Fig. 13, where the iso-density contours and the distribution of GCs are overlapped on to the FDS  $G$ -band image of the region whose limits are  $RA \in (54.146, 54.440)$  and  $Dec.$

**Table 8.** Intersection between our prediction and GCs identified by D’Abrusco et al. (2016) and Cantiello et al. (2018b). Their prediction has been intersected with our set of spectroscopic confirmed GCs (column TRAIN) and with our predicted GCs (column PRED). The percentage refers to the amount of training and predicted sources, i.e. common sources divided by training (or predicted) GCs. The total number of common GCs, together with the fraction (i.e. common GCs divided by all GCs in our work), is reported in column TRAIN + TEST. Columns BLUE and RED are the common blue and red GCs; the percentage refers to the amount of blue and red GCs predicted in this work. The last column of the table (FP) indicates the intersection with the false positives (sources excluded from our prediction), i.e. the number of our removed GCs that D’Abrusco et al. (2016) and Cantiello et al. (2018b) classified as GCs; the percentage refers to the amount of GCs labelled as FPs.

	Common sources identified as GCs in our work					
	TRAIN	PRED	TRAIN + PRED	BLUE	RED	FP
D’Abrusco et al.	329	322	651	322	227	2
(per cent)	92.2	88.9	90.5	72.2	83.2	2.3
Cantiello et al.	286	291	577	358	180	29
(per cent)	80.1	80.4	80.3	80.3	66.0	25.7



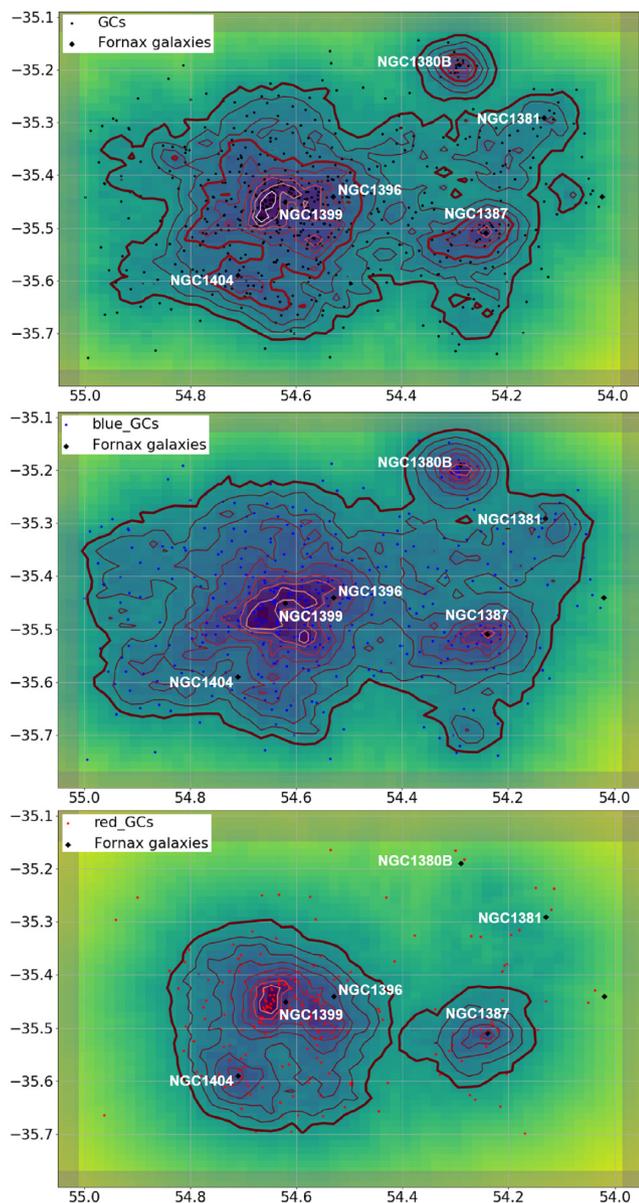
**Figure 11.** CDFs related to colours  $G-I$  (first row) and  $U-I$  (second row) for the whole GC population (first column), for the blue GC subpopulation (second column), and for the red GC subpopulation (third column). In all panels, our CDFs are in blue, D’Abrusco et al. (2016) CDFs are in orange and Cantiello et al. (2018b) CDFs are in red.

$\in (-35.380, -35.620)$ . In this figure is shown the connection between *NGC 1399* and *NGC 1387* which reflects the ‘bridge-like’ stellar stream between the two galaxies made by several filamentary structures, detected by Iodice et al. (2016), whose existence had initially been proposed by Bassino et al. (2006) and confirmed by D’Abrusco et al. (2016). This low surface brightness structure seems to be confirmed by the GCs distribution, whose iso-density contour forms a connection between the two galaxies. This suggests an ongoing interaction between the two galaxies where a fraction of GCs, originally belonging to *NGC 1387*, may have been stripped by the more massive *NGC 1399*.

This interaction between the central structure (*NGC 1399–NGC 1404–NGC 1396*) and *NGC 1387* is particularly evident in the density map of red GCs (bottom panel in Fig. 12). Although *NGC 1396* is a dwarf galaxy separated by  $500 \text{ km s}^{-1}$  from the central galaxies, thus inducing a projection effect, it has been included in analogy with D’Abrusco et al. 2016. Most of the red GCs are

concentrated in two regions, E (75 per cent) and F (10 per cent). The inner contours of *NGC 1399* are characterized by a tail that stretches towards east (i.e. towards *NGC 1387*), maybe due to the interaction between the ellipticals, although it could be a projection effect. Indeed, unlike isolated systems (e.g. the blue GC distribution around *NGC 1380B*), the shape of the contours is strongly irregular, although the incompleteness of GC detections in the centre of giant galaxies could contribute to the observed irregular density structures. The blue GCs density (middle panel in Fig. 12) shows a large and stretched complex region (D, 90 per cent), which connects *NGC 1381* with the mean central overdensity. This suggests that the GC stripping is not confined to *NGC 1387*, but acts on a broader scale.

The fact that no overdensity connected to *NGC 1379*, located at  $(54.02, -35.44)$ , is detected is likely due to the fact that the field covered in this work is limited to the range  $(54.02, 55.38)$ , and the density is underestimated in direction of *NGC 1379*.

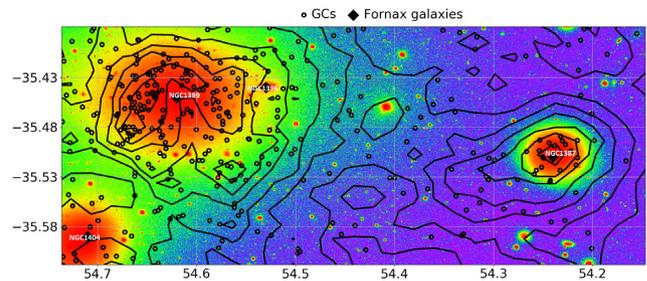


**Figure 12.** Density maps of GC spatial distribution. Top panel: Whole GC population. Middle panel: Blue GC sub-population. Bottom panel: Red GC sub-population. The contours indicate 10 (9 for the sub-populations) log-spaced density levels. The main Fornax galaxies are marked with black filled diamonds.

## 7 CONCLUSIONS

In this work, we have presented an approach based on Astroinformatics methodologies to the identification of GCs from ground-based data. The models under investigation were the GNG, fully implemented using the GPU-oriented Theano library (The Theano Development Team et al. 2016), and the  $\Phi$ Lab FS method (Brescia et al. 2019).

(i) The difference with the standard GNG model is the *batch* sample extraction, which not only allows a faster convergence towards the minimum of the cost function, but also improves scalability, together with the capability to fully exploit the computing resources of the host machine.



**Figure 13.** Detail of the density map overlapped to the FDS  $G$  band in the region that includes *NGC 1399*, *NGC 1396*, *NGC 1404*, *NGC 1387*, with limits  $RA \in (54.146, 54.44)$  and  $Dec. \in (-35.38, -35.62)$ .

(ii) We probed the efficiency of the FS method  $\Phi$ Lab to individuate the complete set of relevant features, by excluding features whose informative contribution was negligible. We have also shown how the relevant set of features found is essentially in agreement with the physics of the problem, since the distribution and projections of the selected hyperspace allow the separation between the class types.

(iii) Comparing the GNG performance with one of the widely used method in Astrophysics, the MLPQNA (Brescia et al. 2012), we confirmed the capability of GNG to separate GCs from background and foreground sources, reaching a satisfying trade-off between purity and completeness, comparable to the MLPQNA results, particularly when the full set of bands was used (i.e. *ugri*).

(iv) Furthermore, we classified an unlabelled set of sources, extracted from the whole catalogue and validated through the limited amount of *HST* detected sources as *ground truth*. Having evidence for candidate false positives, we have applied a GMM in order to exclude them. The bimodal bivariate Gaussian fit returned a set of parameters fully comparable with the literature, thus validating our results.

In order to investigate the prediction capabilities of our methods, the model performances and the identified set of GCs have been compared with other similar works:

(i) By comparing our multiband ground-based results with those obtained with the single-band *HST* photometry (see Brescia et al. 2012; Angora et al. 2017), we showed that, only using all ground-based photometry, the classifiers reach levels of accuracy comparable with those obtained with *HST* single-band photometry. In particular, by introducing the information carried by  $u$  band we reached comparable results to *HST* experiments, although the different efficiencies of the instruments limit the ground-based analysis to brighter sources.

(ii) The matching with the results obtained by D’Abrusco et al. (2016) and Cantiello et al. (2018b) probed the robustness of our method, fully comparable with other techniques which exploit different approaches.

(iii) Finally, the density maps for the red, blue and whole GC populations showed the usefulness of our prediction method and underlined some interesting features of the Fornax core, comparable to other studies (Bassino et al. 2006; D’Abrusco et al. 2016; Iodice et al. 2016; Cantiello et al. 2018b).

Although our approach requires a spectroscopic knowledge in order to build a broad and pure KB, indispensable for training ML models, the method avoids the introduction of arbitrary photometric cuts, which, although plausible, are bound to a maximum of three-dimensional viewpoints of the phenomenology, thus unavoidably

originating contamination effects. It is important to underline that our results are constrained by the exiguous number of labelled sources available in the catalogue and by the unavoidable inhomogeneities among the filters, particularly concerning the presence of  $\sim 82$  per cent missing data among the *u*-band samples, which we proved to be crucial to effectively separate the GCs from different types of sources. In future works, an improvement could be obtained by a reduction of VST data tailored to compact sources (in progress) or by including external photometry such as, e.g. DECam *u* band (Abbott et al. 2018). The ongoing reduction of the full FDS survey data, will allow to extend these results to the whole Fornax cluster out to the virial radius.

## ACKNOWLEDGEMENTS

The authors thank the anonymous referee for all very helpful comments and suggestions that improved the scientific quality of the presented work. MB acknowledges the *INAF Progetto di Ricerca di Interesse Nazionale - Square Kilometer Array (PRIN-SKA) 2017 program 1.05.01.88.04* and the funding from *MIUR Premiale 2016: Mining the Cosmos - Big Data and Innovative Italian technology for Frontier Astrophysics and Cosmology (MITIC)*. MP acknowledges support from *Progetto di Ricerca di Interesse Nazionale (PRIN) INAF 2014 'Fornax Cluster Imaging and Spectroscopic Deep Survey'*. MP and SC acknowledge support from the project 'Quasars at high redshift: physics and cosmology' financed by the ASI/INAF agreement 2017-14-H.0. GL, RP, and NN acknowledge support from the European Union's Horizon 2020 Sundial Innovative Training Network, grant no. 721463. NN acknowledges support from the 100 Top Talent Program of the Sun Yat-sen University, Guangdong Province. MS and EI acknowledge financial support from the VST project. RD'A is supported by NASA contract NAS8-03060 (Chandra X-ray Center). GD acknowledges support from *Comisión Nacional de Investigación Científica y Tecnológica (CONICYT) project Basal AFB-170002. DAMEWARE* has been used for ML experiments (Brescia et al. 2014). Topcat has been used for this work (Taylor 2005). C<sup>3</sup> has been used for efficient catalogue cross-matching (Riccio et al. 2017).

## REFERENCES

- Abbott T. M. C. et al., 2018, *ApJS*, 239, 18
- Angora G., Brescia M., Riccio G., Cavuoti S., Paolillo M., Puzia T. H., 2017, in Kalinichenko L., Manolopoulos Y., Skvortsov N., Sukhomlin V., eds, Selected Papers of the XIX International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2017), CEUR Workshop Proceedings, Vol. 2022, Kalinichenko L., Moscow, Russia, p. 81
- Ashman K., Zepf S., 2008, *Globular Cluster Systems*. Cambridge Astrophysics, Cambridge Univ. Press, Cambridge
- Ashman K. M., Bird C. M., Zepf S. E., 1994, *AJ*, 108, 2348
- Baron D., 2019, preprint (arXiv:1904.07248)
- Bassino L. P., Faifer F. R., Forte J. C., Dirsch B., Richtler T., Geisler D., Schuberth Y., 2006, *A&A*, 451, 789
- Batista G. E. A. P. A., Monard M. C., 2003, *Appl. Artif. Intell.*, 17, 519
- Bertin E., Arnouts S., 1996, *A&AS*, 117, 393
- Bishop C. M., 2006, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA
- Borne K. et al., 2009, in *astro2010: The Astronomy and Astrophysics Decadal Survey*, The National Academies Press, Washington, DC
- Bortoletti A., Di Fiore C., Fanelli S., Zellini P., 2003, *IEEE Trans. Neural Netw.*, 14, 263
- Breiman L., 2001, *Mach. Learn.*, 45, 5
- Breiman L., Last M., Rice J., 2003, in *Statistical Challenges in Astronomy*. Springer, New York, NY, p. 243
- Brescia M., Longo G., 2013, *Nucl. Instrum. Methods Phys. Res. A*, 720, 92
- Brescia M., Cavuoti S., Paolillo M., Longo G., Puzia T., 2012, *MNRAS*, 421, 1155
- Brescia M., Cavuoti S., D'Abrusco R., Longo G., Mercurio A., 2013, *ApJ*, 772, 140
- Brescia M. et al., 2014, *PASP*, 126, 783
- Brescia M., Cavuoti S., Amaro V., Riccio G., Angora G., Vellucci C., Longo G., 2018, *Commun. Comput. Inform. Sci.*, 822, 61
- Brescia M., Salvato M., Cavuoti S., Ananna T. T., Riccio G., LaMassa S. M., Urry C. M., Longo G., 2019, *MNRAS*, 489, 663
- Brodie J. P., Strader J., 2006, *ARA&A*, 44, 193
- Byrd R., Nocedal J., Schnabel R., 1994, *Math. Program.*, 63, 129
- Camino R. D., Hammerschmidt C. A., State R., 2019, Improving Missing Data Imputation with Deep Generative Models, CoRR, preprint (arXiv:1902.10666)
- Cantiello M., Grado A., Rejkuba M., Arnaboldi M., Capaccioli M., Greggio L., Iodice E., Limatola L., 2018a, *A&A*, 611, A21
- Cantiello M. et al., 2018b, *A&A*, 611, A93
- Cavuoti S., Brescia M., Longo G., Mercurio A., 2012, *A&A*, 546, A13
- Cavuoti S., Garofalo M., Brescia M., Pescape A., Longo G., Ventre G., 2013, *Smart Innov. Syst. Technol.*, 19, 29
- Cavuoti S., Brescia M., De Stefano V., Longo G., 2015, *Exp. Astron.*, 39, 45
- D'Abrusco R., Fabbiano G., Zezas A., 2015, *ApJ*, 805, 26
- D'Abrusco R. et al., 2016, *ApJ*, 819, L31
- Delli Veneri M., Cavuoti S., Brescia M., Longo G., Riccio G., 2019, *MNRAS*, 486, 1377
- D'Isanto A., Cavuoti S., Brescia M., Donalek C., Longo G., Riccio G., Djorgovski S., 2016, *MNRAS*, 457, 3119
- Duda R. O., Hart P. E., Stork D. G., 2000, *Pattern Classification*, 2nd edn. Wiley-Interscience, Hoboken, New Jersey
- Fasano G., Franceschini A., 1987, *MNRAS*, 225, 155
- Fawcett T., 2006, *Pattern Recognit. Lett.*, 27, 861
- Feigelson E., Hilbe J. M., 2014, *American Astronomical Society Meeting Abstracts #223*, p. 253.02
- Floudas C. C. A., Pardalos P. M., 2006, *Encyclopedia of Optimization*. Springer-Verlag, Berlin, Heidelberg
- Fritzke B., 1994, *Neural Netw.*, 7, 1441
- Fritzke B., 1995, in Tesauro G., Touretzky D. S., Leen T. K., eds, *Advances in Neural Information Processing Systems 7*. MIT Press, Cambridge, Massachusetts, p. 625
- Geisler D., Forte J. C., 1990, *ApJ*, 350, L5
- Gheyas I. A., Smith L. S., 2010, *Pattern Recognit.*, 43, 5
- Guyon I., Elisseeff A., 2003, *J. Mach. Learn. Res.*, 3, 1157
- Guyon I., Gunn S., Nikravesh M., Zadeh L. A., 2006, *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*. Springer-Verlag, Berlin, Heidelberg
- Hanley J. A., McNeil B. J., 1982, *Radiology*, 143, 29
- Hara S., Maehara T., 2016, preprint (arXiv:1611.05940)
- Hara S., Maehara T., 2017, in Singh S., Markovitch S., eds, 31st AAAI Conference on Artificial Intelligence, AAAI 2017, AAAI Press, San Francisco, California, p. 1985
- Hastie T., Tibshirani R., Friedman J., 2001, *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA
- Hastie T., Tibshirani R., Friedman J., 2009, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics, 2nd edn. Springer, New York
- Iodice E. et al., 2016, *ApJ*, 820, 42
- Iodice E. et al., 2017, *ApJ*, 839, 21
- Jain A., Zongker D., 1997, *IEEE Trans. Pattern Anal. Mach. Intell.*, 19, 153
- Jolliffe I. T., 2002, *Principal Component Analysis*. Springer, New York
- Kohavi R., 1995, in Proceedings of the 14th International Joint Conference on Artificial Intelligence, IJCAI'95, Vol. 2. Morgan Kaufmann Publishers Inc., San Francisco, CA, p. 1137

- Kohavi R., John G. H., 1997, *Artif. Intell.*, 97, 273
- Kuijken K., 2011, *The Messenger*, 146, 8
- Kullback S., Leibler R. A., 1951, *Ann. Math. Stat.*, 22, 79
- Kundu A., Whitmore B. C., 1998, *AJ*, 116, 2841
- Kursa M., Rudnicki W., 2010, *J. Stat. Softw.*, 36, 1
- Lal T., Chapelle O., Weston J., Elisseeff A., 2006, *Embedded Methods*. Springer, Berlin, Germany, p. 137
- Marlin B., 2008, PhD thesis. Department of Computer Science, Univ. Toronto
- Martinetz T. M., Schulten K. J., 1991, in Kohonen T., Mäksäsa K., Simula O., Kangas J., eds, *Proceedings of the International Conference on Artificial Neural Networks 1991*. North-Holland, Amsterdam, New York, p. 397
- Martinetz T. M., Berkovich S. G., Schulten K. J., 1993, *Trans. Neural Netw.*, 4, 558
- Montoro J. C. G., Abascal J. L. F., 1993, *J. Phys. Chem.*, 97, 4211
- Muñoz R. P. et al., 2014, *ApJS*, 210, 4
- Muratov A. L., Gnedin O. Y., 2010, *ApJ*, 718, 1266
- Murtagh F., Legendre P., 2014, *J. Classif.*, 31, 274
- Nakoneczny S., Bilicki M., Solarz A., Pollo A., Maddox N., Spiniello C., Brescia M., Napolitano N. R., 2019, *A&A*, 624, A13
- Parker R., 2010, *Missing Data Problems in Machine Learning*. VDM Verlag, Riga, Latvia
- Peacock J. A., 1983, *MNRAS*, 202, 615
- Pedregosa F. et al., 2011, *J. Mach. Learn. Res.*, 12, 2825
- Pota V. et al., 2013, *MNRAS*, 428, 389
- Pota V. et al., 2018, *MNRAS*, 481, 1744
- Poulos J., Valle R., 2016, preprint ([arXiv:1610.09075](https://arxiv.org/abs/1610.09075))
- Puzia T. H., Paolillo M., Goudfrooij P., Maccarone T. J., Fabbiano G., Angelini L., 2014, *ApJ*, 786, 78
- Riccio G., Brescia M., Cavuoti S., Mercurio A., di Giorgio A. M., Molinari S., 2017, *PASP*, 129, 024005
- Russell S. J., Norvig P., 2010, *Artificial Intelligence*, Prentice Hall Series in Artificial Intelligence, 3rd edn. Prentice Hall, Upper Saddle River
- Schipani P. et al., 2012, *Mem. Soc. Astron. Ital. Suppl.*, 19, 393
- Schuberth Y., Richtler T., Hilker M., Dirsch B., Bassino L. P., Romanowsky A. J., Infante L., 2010, *A&A*, 513, A52
- Stehman S. V., 1997, *Remote Sens. Environ.*, 62, 77
- Tangaro S. et al., 2015, *Comput. Math. Methods Med.*, 2015, 10
- Taylor J. R., 1996, *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*, 2 sub edn. University Science Books, Mill Valley, CA
- Taylor M. B., 2005, in Shopbell P., Britton M., Ebert R., eds, *ASP Conf. Ser. Vol. 347, Astronomical Data Analysis Software and Systems XIV*. Astron. Soc. Pac., San Francisco, p. 29
- The Theano Development Team et al., 2016, preprint ([arXiv:1605.02688](https://arxiv.org/abs/1605.02688))
- Tibshirani R. J., 2013, *Electron. J. Stat.*, 7, 1456
- Tikhonov A., 1998, *Nonlinear Ill-Posed Problems*. Springer, Netherlands
- Van Der Maaten L., 2014, *J. Mach. Learn. Res.*, 15, 3221
- van der Maaten L., Hinton G., 2008, *J. Mach. Learn. Res.*, 9, 2579
- Wittmann C., Lisker T., Pasquali A., Hilker M., Grebel E. K., 2016, *MNRAS*, 459, 4450
- Yoon J., Jordon J., van der Schaar M., 2018, preprint ([arXiv:1806.02920](https://arxiv.org/abs/1806.02920))
- Zepf S. E., Ashman K. M., Geisler D., 1995, *ApJ*, 443, 570
- Zhang H., Xie P., Xing E., 2018, preprint ([arXiv:1808.01684](https://arxiv.org/abs/1808.01684))

## APPENDIX A: FEATURE IMPORTANCES

In this section, we report five tables regarding the FS process. Table A1 shows the importance values estimated by  $\Phi$ LAB to select the PS used in the six classification experiments. The top panel encloses the selected features, while bottom panel contains the rejected ones. Table A2 specifies the involved features for the four experiments chosen to validate the FS performed by  $\Phi$ LAB (see Section 4.1). The results achieved by the GNG and RF are, respectively, shown in Tables A4 and A6. Finally, Table A5 illustrates the results achieved by the GNG and RF on a data set composed by: (i) the same *ugri* samples after having removed the information regarding the *u* band; (ii) a data set whose features represent only the information carried by the *u* band, (iii) the whole *ugri* informative contribution (taken from Tables 1 and A6).

**Table A1.** Feature importance values for the features selected by  $\Phi$ LAB (top table) and for the features rejected (bottom table), related to the six performed experiments.

SELECTED FEATURE	3CLASS	<i>ugri</i> GCALL	GCSTAR	3CLASS	<i>gri</i> GCALL	GCSTAR
<i>u</i> FWHM	0.0261	0.0085	0.0094			
<i>g</i> FWHM	0.0811	0.0281	0.0240	0.0639	0.0294	0.0203
<i>r</i> FWHM	0.0676	0.0407	0.0221	0.0848	0.0272	0.0257
<i>i</i> FWHM	0.0924	0.0225	0.0181	0.0609	0.0257	0.0391
<i>u</i> FLUX RADIUS	0.0290	0.0084	0.0060			
<i>g</i> FLUX RADIUS	0.0431	0.0321	0.0120	0.0782	0.0231	0.0139
<i>r</i> FLUX RADIUS	0.0349	0.0188	0.0137	0.0485	0.0312	0.0098
<i>i</i> FLUX RADIUS	0.0390	0.0210	0.0069	0.0804	0.0162	0.0071
<i>u</i> MAG AUTO	0.0106	0.0320	0.0191			
<i>g</i> MAG APER4	0.0057	0.0130	0.0174			
<i>r</i> MAG APER6	0.0107	0.0080	0.0068			
<i>i</i> MAG APER8	0.0063	0.0087	0.0149			
<i>g</i> MAG AUTO	0.0108	0.0115	0.0145	0.0120	0.0105	0.0150
<i>g</i> MAG APER4	0.0199	0.0046	0.0132	0.0090	0.0082	0.0162
<i>g</i> MAG APER6	0.0188	0.0111	0.0179	0.0101	0.0075	0.0118
<i>g</i> MAG APER8	0.0066	0.0051	0.0245	0.0089	0.0128	0.0212
<i>r</i> MAG AUTO	0.0127	0.0247	0.0205	0.0190	0.0184	0.0213
<i>r</i> MAG APER4	0.0148	0.0107	0.0600	0.0124	0.0137	0.0144
<i>r</i> MAG APER6	0.0195	0.0106	0.0432	0.0083	0.0124	0.0232
<i>r</i> MAG APER8	0.0163	0.0145	0.0359	0.0089	0.0167	0.0228
<i>i</i> MAG AUTO	0.0058	0.0099	0.0081	0.0155	0.0274	0.0373

**Table A1** – *continued*

SELECTED FEATURE	3CLASS	<i>ugri</i> GCALL	GCSTAR	3CLASS	<i>gri</i> GCALL	GCSTAR
<i>i</i> MAG APER4	0.0087	0.0131	0.0231	0.0149	0.0211	0.0469
<i>i</i> MAG APER6	0.0063	0.0123	0.0200	0.0190	0.0208	0.0414
<i>i</i> MAG APER8	0.0072	0.0100	0.0148	0.0176	0.0162	0.0399
<i>u</i> – <i>g</i> AUTO	0.0092	0.0244	0.0120			
<i>u</i> – <i>g</i> APER4	0.0301	0.0220	0.0096			
<i>u</i> – <i>g</i> APER6	0.0120	0.0260	0.0103			
<i>u</i> – <i>g</i> APER8	0.0132	0.0393	0.0109			
<i>g</i> – <i>r</i> AUTO	0.0087	0.0183	0.0108	0.0173	0.0350	0.0302
<i>g</i> – <i>r</i> APER4	0.0054	0.0130	0.0100	0.0154	0.0239	0.0299
<i>g</i> – <i>r</i> APER6	0.0074	0.0185	0.0090	0.0160	0.0302	0.0283
<i>g</i> – <i>r</i> APER8	0.0094	0.0185	0.0108	0.0146	0.0263	0.0452
<i>r</i> – <i>i</i> AUTO	0.0047	0.0065	0.0078	0.0083	0.0142	0.0150
<i>r</i> – <i>i</i> APER4	0.0057	0.0067	0.0131	0.0130	0.0154	0.0198
<i>r</i> – <i>i</i> APER6	0.0065	0.0059	0.0107	0.0135	0.0171	0.0191
<i>r</i> – <i>i</i> APER8	0.0053	0.0074	0.0092	0.0089	0.0198	0.0235
<i>u</i> MU MAX	0.0090	0.0085	0.0195			
<i>g</i> MU MAX	0.0190	0.0118	0.0383	0.0121	0.0110	0.0179
<i>r</i> MU MAX	0.0206	0.0082	0.0683	0.0152	0.0127	0.0222
<i>i</i> MU MAX	0.0120	0.0102	0.0145	0.0178	0.0146	0.0552
<i>u</i> A WORLD	0.0085	0.0327	0.0037			
<i>g</i> A WORLD	0.0105	0.0287	0.0044	0.0148	0.0492	0.0058
<i>r</i> A WORLD	0.0129	0.0248	0.0047	0.0190	0.0292	0.0050
<i>i</i> A WORLD	0.0116	0.0247	0.0052	0.0258	0.0420	0.0081
<i>u</i> B WORLD	0.0043	0.0206	0.0079			
<i>g</i> B WORLD	0.0088	0.0263	0.0048	0.0273	0.0449	0.0109
<i>r</i> B WORLD	0.0084	0.0217	0.0046	0.0207	0.0492	0.0094
<i>i</i> B WORLD	0.0082	0.0360	0.0062	0.0227	0.0499	0.0298
<i>i</i> PETRO RADIUS	0.0112	0.0021	0.0042	0.0069	0.0033	0.0090
REJECTED FEATURE	3CLASS	<i>ugri</i> GCALL	GCSTAR	3CLASS	<i>gri</i> GCALL	GCSTAR
<i>u</i> PETRO RADIUS	0.0018	0.0027	0.0012			
<i>g</i> PETRO RADIUS	0.0156	0.0011	0.0014	0.0086	0.0030	0.0034
<i>r</i> PETRO RADIUS	0.0084	0.0012	0.0054	0.0063	0.0032	0.0046
<i>u</i> KRON RADIUS	0.0028	0.0034	0.0051			
<i>g</i> KRON RADIUS	0.0002	0.0002	0.0001	0.0004	0.0010	0.0006
<i>r</i> KRON RADIUS	0.0003	0.0003	0.0001	0.0009	0.0010	0.0009
<i>i</i> KRON RADIUS	0.0006	0.0005	0.0006	0.0014	0.0025	0.0023
<i>u</i> ELONG	0.0032	0.0047	0.0027			
<i>g</i> ELONG	0.0038	0.0028	0.0031	0.0049	0.0049	0.0038
<i>r</i> ELONG	0.0029	0.0034	0.0048	0.0058	0.0050	0.0047
<i>i</i> ELONG	0.0024	0.0029	0.0044	0.0050	0.0049	0.0059
<i>u</i> THETA	0.0018	0.0029	0.0025			
<i>g</i> THETA	0.0013	0.0034	0.0041	0.0023	0.0040	0.0041
<i>r</i> THETA	0.0017	0.0026	0.0024	0.0021	0.0041	0.0041
<i>i</i> THETA	0.0018	0.0028	0.0030	0.0025	0.0036	0.0043

**Table A2.** List of features composing the four PSs used for validation of the FS process with the method  $\Phi$ LAB: column FULL refers to the original PS, including all available features; column BEST refers to the best solution obtained by the FS method; column MIXED refers to a variant of the BEST PS obtained by replacing all 15 features rejected by  $\Phi$ LAB to a subset randomly extracted from the best solution; finally, column BEST + REJECTED is another variant of the BEST PS, where the 15 rejected features were inserted in place of the least significant features of the BEST PS.

FEATURE	FULL	BEST	MIXED	BEST + REJECTED
<i>u</i> FWHM	×	×	×	×
<i>g</i> FWHM	×	×		×
<i>r</i> FWHM	×	×	×	×
<i>i</i> FWHM	×	×	×	×
<i>u</i> FLUX RADIUS	×	×		×
<i>g</i> FLUX RADIUS	×	×	×	×
<i>r</i> FLUX RADIUS	×	×	×	×
<i>i</i> FLUX RADIUS	×	×	×	
<i>u</i> MAG AUTO	×	×		×
<i>u</i> MAG APER4	×	×	×	×
<i>u</i> MAG APER6	×	×		×
<i>u</i> MAG APER8	×	×		×
<i>g</i> MAG AUTO	×	×	×	
<i>g</i> MAG APER4	×	×	×	×
<i>g</i> MAG APER6	×	×		
<i>g</i> MAG APER8	×	×	×	
<i>r</i> MAG AUTO	×	×		×
<i>r</i> MAG APER4	×	×	×	
<i>r</i> MAG APER6	×	×	×	×
<i>r</i> MAG APER8	×	×	×	×
<i>i</i> MAG AUTO	×	×	×	
<i>i</i> MAG APER4	×	×	×	×
<i>i</i> MAG APER6	×	×	×	×
<i>i</i> MAG APER8	×	×	×	
<i>u-g</i> AUTO	×	×	×	×
<i>u-g</i> APER4	×	×		×
<i>u-g</i> APER6	×	×	×	×
<i>u-g</i> APER8	×	×		×
<i>g-r</i> AUTO	×	×	×	×
<i>g-r</i> APER4	×	×	×	×
<i>g-r</i> APER6	×	×	×	×
<i>g-r</i> APER8	×	×		×
<i>r-i</i> AUTO	×	×	×	
<i>r-i</i> APER4	×	×		
<i>r-i</i> APER6	×	×	×	×
<i>r-i</i> APER8	×	×	×	
<i>u</i> MU MAX	×	×		×
<i>g</i> MU MAX	×	×	×	
<i>r</i> MU MAX	×	×	×	×
<i>i</i> MU MAX	×	×		×
<i>u</i> A WORLD	×	×	×	
<i>g</i> A WORLD	×	×		×
<i>r</i> A WORLD	×	×	×	×
<i>i</i> A WORLD	×	×	×	×
<i>u</i> B WORLD	×	×	×	×
<i>g</i> B WORLD	×	×	×	
<i>r</i> B WORLD	×	×	×	
<i>i</i> B WORLD	×	×		×
<i>i</i> PETRO RADIUS	×	×	×	
<i>u</i> PETRO RADIUS	×		×	×
<i>g</i> PETRO RADIUS	×		×	×
<i>r</i> PETRO RADIUS	×		×	×
<i>u</i> KRON RADIUS	×		×	×
<i>g</i> KRON RADIUS	×		×	×
<i>r</i> KRON RADIUS	×		×	×
<i>i</i> KRON RADIUS	×		×	×
<i>u</i> ELONG	×		×	×
<i>g</i> ELONG	×		×	×
<i>r</i> ELONG	×		×	×
<i>i</i> ELONG	×		×	×

**Table A2** – *continued*

FEATURE	FULL	BEST	MIXED	BEST + REJECTED
<i>u</i> THETA	×		×	×
<i>g</i> THETA	×		×	×
<i>r</i> THETA	×		×	×
<i>i</i> THETA	×		×	×
<b>TOTAL</b>	<b>64</b>	<b>49</b>	<b>49</b>	<b>49</b>

**Table A3.** Classification results in terms of statistical estimators (the same used in Table 1) achieved by GNG and *K*-means on *ugri* and *gri* data sets, using the *BEST* PS. The results refer to the three described classification problems: *3CLASS* (top table), *GCs versus ALL* (middle table), *GCs versus STARS* (bottom table).

3CLASS		<i>ugri</i>		<i>gri</i>	
ESTIMATOR (per cent)	GNG	<i>K</i> -means	GNG	<i>K</i> -means	
AE	85.5	82.9	79.4	82.1	
pur STAR	89.4	81.3	77.9	79.9	
compl STAR	76.8	70.6	65.4	62.1	
F1 STAR	83.1	76.0	71.6	71.0	
pur GCs	77.7	70.4	75.3	75.8	
compl GCs	86.9	80.6	82.0	81.6	
F1 GCs	82.3	75.5	78.5	78.7	
pur gal	93.5	93.6	88.9	89.0	
compl gal	92.1	92.0	91.1	90.6	
F1 gal	92.8	92.8	90.0	89.8	
GCs versus ALL		<i>ugri</i>		<i>gri</i>	
ESTIMATOR (per cent)	GNG	<i>K</i> -means	GNG	<i>K</i> -means	
AE	87.8	86.0	82.8	80.4	
pur notGC	82.2	75.4	81.1	78.5	
compl notGC	89.6	83.5	82.8	75.2	
F1 notGC	85.9	79.4	81.9	76.9	
pur GCs	91.5	90.8	87.1	81.7	
compl GCs	87.4	87.2	80.3	80.3	
F1 GCs	89.4	89.0	83.7	81.0	
GCs versus STARS		<i>ugri</i>		<i>gri</i>	
ESTIMATOR (per cent)	GNG	<i>K</i> -means	GNG	<i>K</i> -means	
AE	83.8	85.7	77.2	80.5	
pur STAR	83.8	87.4	75.5	81.6	
compl STAR	88.3	93.6	82.9	90.8	
F1 STAR	86.0	90.5	79.2	86.2	
pur GCs	88.3	80.0	81.3	81.4	
compl GCs	81.0	65.7	71.0	59.4	
F1 GCs	84.6	72.8	76.1	70.4	

**Table A4.** GNG classification results in terms of statistical estimators (the same used in Table 1) for both *ugri* and *gri* data set types. Top table reports the results for the 3-class experiment, middle table for the 2-class experiment between GCs and *not*GCs (stars + galaxies), and bottom table shows the results concerning the 2-class experiment between GCs and stars. The columns BEST, FULL, MIXED, and BEST + REJECTED are related to the four PSs, described in Table A2.

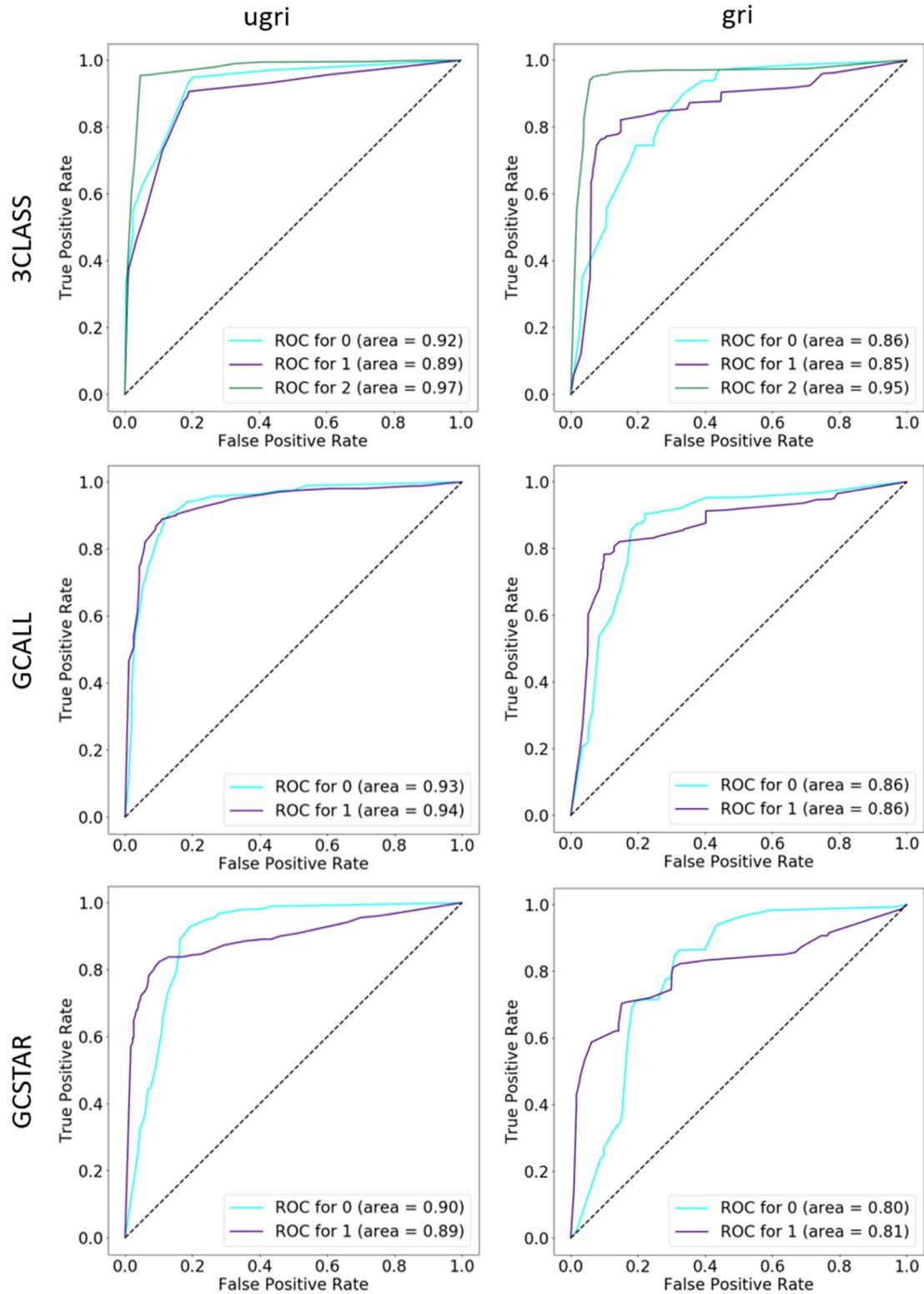
3CLASS		<i>ugri</i>			<i>gri</i>			
ESTIMATOR (per cent)	BEST	FULL	MIXED	BEST + REJECTED	BEST	FULL	MIXED	BEST + REJECTED
AE	86.5	51.9	40.0	55.3	79.4	55.8	30.8	30.0
pur STAR	85.8	33.3	37.7	39.0	71.9	29.4	0	0
compl STAR	80.3	9.7	7.3	8.8	66.9	28.1	0	0
F1 STAR	83.0	15.0	12.2	14.4	69.3	28.8	0	0
pur GCs	80.0	42.3	43.0	45.0	78.2	51.7	43.2	43.2
compl GCs	90.8	31.7	37.5	79.6	79.6	53.2	56.8	50.2
F1 GCs	85.1	36.2	40.0	57.5	78.9	52.4	49.1	46.5
pur gal	92.5	44.3	44.5	44.6	88.3	56.1	28.3	28.3
compl gal	95.4	78.5	78.4	80.8	92.1	56.1	50.0	49.9
F1 gal	93.9	56.9	56.7	57.4	90.2	56.1	36.1	36.1
GCs versus ALL		<i>ugri</i>			<i>gri</i>			
ESTIMATOR (per cent)	BEST	FULL	MIXED	BEST + REJECTED	BEST	FULL	MIXED	BEST + REJECTED
AE	88.7	59.6	59.1	59.3	84.0	60.0	66.4	66.3
pur notGC	85.1	51.3	50.7	51.1	81.3	62.4	71.4	71.4
compl notGC	88.0	44.0	40.3	40.6	85.5	42.3	48.9	48.9
F1 notGC	86.5	47.4	45.0	45.2	83.4	52.3	60.1	58.1
pur GCs	91.3	64.1	63.2	63.4	86.2	60.3	63.9	63.9
compl GCs	89.1	70.5	72.3	72.5	82.1	76.2	82.2	82.2
F1 GCs	90.2	67.2	67.5	67.7	84.2	68.2	71.9	71.9
GCs versus STARS		<i>ugri</i>			<i>gri</i>			
ESTIMATOR (per cent)	BEST	FULL	MIXED	BEST + REJECTED	BEST	FULL	MIXED	BEST + REJECTED
AE	86.8	55.0	52.9	50.8	78.2	50.0	50.0	50.0
pur STAR	87.0	56.5	54.5	52.8	77.3	54.2	54.2	54.2
compl STAR	83.2	52.1	49.3	47.1	84.9	50.1	50.0	49.6
F1 STAR	85.1	54.2	51.7	49.5	81.1	52.1	52.1	51.9
pur GCs	91.6	53.6	51.6	49.6	79.7	45.6	45.6	45.4
compl GCs	80.3	57.9	56.8	54.7	70.3	49.8	49.8	49.4
F1 GCs	85.6	55.7	54.1	52.0	75.0	47.6	47.6	47.4

**Table A5.** Classification results in terms of statistical estimators (the same used in Table 1) achieved by GNG and RF on different data sets: (i) composed only by the information carried by the  $u$  band (first two columns), (ii) composed by same  $ugri$  samples without the information carried by the  $u$  band (named as  $gri^*$ , third and fourth columns), (iii) the whole  $ugri$  informative contribution (last two columns, the performances are taken from Tables 1 and A6). The results refer to the three described classification problems: *3CLASS* (top table), *GCS versus ALL* (middle table), *GCS versus STARS* (bottom table).

3CLASS		$u$		$gri^*$		$ugri$	
ESTIMATOR (per cent)	GNG	RF	GNG	RF	GNG	RF	
AE	85.5	84.1	79.4	82.4	86.5	94.4	
pur STAR	89.4	80.4	77.9	81.1	85.8	85.4	
compl STAR	76.8	81.7	65.4	79.8	80.3	85.5	
F1 STAR	83.1	81.0	71.6	80.4	83.0	85.5	
pur GCS	77.7	79.3	75.3	80.3	80.0	86.9	
compl GCS	86.9	80.8	82.0	80.2	90.8	93.3	
F1 GCS	82.3	80.1	78.5	80.2	85.1	90.0	
pur gal	93.5	93.3	88.9	90.7	92.5	95.6	
compl gal	92.1	90.1	91.1	85.2	95.4	97.7	
F1 gal	92.8	91.7	90.0	87.9	93.9	96.6	
GCS versus ALL		$u$		$gri^*$		$ugri$	
ESTIMATOR (per cent)	GNG	RF	GNG	RF	GNG	RF	
AE	87.8	86.6	82.8	85.0	88.7	92.2	
pur notGC	82.2	81.2	81.1	80.2	85.1	91.0	
compl notGC	89.6	80.8	82.8	83.8	88.0	89.9	
F1 notGC	85.9	81.0	81.9	82.0	86.5	90.4	
pur GCS	91.5	90.8	87.1	88.0	91.3	92.8	
compl GCS	87.4	90.1	80.3	84.1	89.1	92.6	
F1 GCS	89.4	90.5	83.7	86.0	90.2	92.7	
GCS versus STARS		$u$		$gri^*$		$ugri$	
ESTIMATOR (per cent)	GNG	RF	GNG	RF	GNG	RF	
AE	83.8	90.8	77.2	87.0	86.8	88.2	
pur STAR	83.8	90.7	75.5	84.8	87.0	85.9	
compl STAR	88.3	97.1	82.9	86.0	83.2	92.2	
F1 STAR	86.0	93.9	79.2	85.4	85.1	88.9	
pur GCS	88.3	92.6	81.3	89.1	91.6	90.8	
compl GCS	81.0	80.4	71.0	80.4	80.3	95.2	
F1 GCS	84.6	86.5	76.1	84.7	85.6	92.9	

**Table A6.** Random Forest classification results in terms of statistical estimators (the same used in Table 1) for both *ugri* and *gri* data set types. Top table reports the results for the 3-class experiment, middle table for the 2-class experiment between GCs and *notGCs* (stars + galaxies), and bottom table shows the results concerning the 2-class experiment between GCs and stars. The columns BEST, FULL, MIXED, and BEST + REJECTED are related to the four PSs, described in Table A2.

3CLASS ESTIMATOR (per cent)	BEST	FULL	<i>ugri</i> MIXED	BEST + REJECTED	BEST	FULL	<i>gri</i> MIXED	BEST + REJECTED
AE	94.4	92.6	93.0	92.7	92.6	91.1	90.2	90.2
pur STAR	85.4	84.6	85.1	84.7	83.7	80.5	77.1	78.9
compl STAR	85.7	86.1	85.7	86.1	84.1	80.6	80.5	80.2
F1 STAR	85.5	85.3	85.4	85.4	83.9	80.5	78.8	79.5
pur GCs	86.9	87.2	88.0	86.9	88.9	86.5	86.0	86.7
compl GCs	93.3	85.5	85.1	85.1	91.2	90.2	88.8	89.6
F1 GCs	90.0	86.3	86.5	86.0	90.0	88.3	87.4	88.1
pur gal	95.6	97.1	96.8	95.8	94.6	94.6	95.1	95.0
compl gal	97.7	97.0	96.1	97.8	96.6	89.9	88.8	89.1
F1 gal	96.6	97.0	96.4	96.8	95.6	92.2	91.5	92.0
GCs versus ALL ESTIMATOR (per cent)	BEST	FULL	<i>ugri</i> MIXED	BEST + REJECTED	BEST	FULL	<i>gri</i> MIXED	BEST + REJECTED
AE	92.2	92.0	91.7	92.3	88.1	87.4	86.1	87.1
pur notGC	91.0	90.1	90.2	90.3	84.6	83.0	83.4	83.2
compl notGC	89.9	90.2	89.7	89.6	91.3	91.3	90.8	91.2
F1 notGC	90.4	90.1	89.9	89.9	87.8	87.0	86.9	87.0
pur GCs	92.8	93.0	92.7	92.7	92.2	91.3	90.7	92.1
compl GCs	92.6	91.1	90.4	92.8	84.9	83.1	82.0	82.6
F1 GCs	92.7	92.0	91.5	92.7	88.4	86.8	86.1	87.1
GCs versus STARS ESTIMATOR (per cent)	BEST	FULL	<i>ugri</i> MIXED	BEST + REJECTED	BEST	FULL	<i>gri</i> MIXED	BEST + REJECTED
AE	88.2	88.0	87.8	87.6	88.1	88.1	87.3	87.7
pur STAR	85.9	85.9	86.3	85.7	85.9	85.7	83.7	85.9
compl STAR	92.2	90.4	91.7	91.6	92.7	93.8	92.8	92.7
F1 STAR	88.9	88.1	88.9	88.6	89.2	88.7	88.0	89.2
pur GCs	90.8	89.7	90.7	91.0	91.1	91.7	90.2	91.1
compl GCs	95.2	84.2	84.1	85.1	83.4	83.9	83.8	81.4
F1 GCs	92.9	86.9	87.3	88.0	87.1	87.6	86.9	86.0



**Figure A1.** ROC curves related to the GNG performances for the six experiments. GCs are labelled as 1 (purple), stars are labelled with 0 (light blue) in the *3CLASS* and *GCSTAR* experiments, while galaxies are labelled with 2 in the *3CLASS* experiments (green), in the *GCALL* experiments the label 0 refers to the *notGCs*. In all the panels is reported the area and the curve and the non-discrimination line (dotted).

This paper has been typeset from a  $\text{\TeX}/\text{\LaTeX}$  file prepared by the author.