The Astrophysical Journal, 1005:96 (22pp), 2026 July 1 https://doi.org/10.3847/1538-4357/ae7321

The Contribution of the Color Space in LSST-like Photometry for the Selection of Extragalactic Globular Cluster Candidates

Nicholas Schweder-Souza[1,2], Ana L. Chies-Santos[2,3], Rafael S. de Souza[3,4,5], Kristen C. Dage[6], Charles J. Bonatto[2,3], Juan P. Caso[7,8], Michele Cantiello[9], Pedro dos Santos-Lopes[2,3], Pedro Floriano[2,3], Thayse A. Pacheco[3,10], Katherine L. Rhode[11], Pauline Barmby[12], Jennifer Sobeck[13], Ana I. Ennis[14,15], Yasna Ordenes-Briceño[16], Teymoor Saifollahi[10,17], Julia Gschwend[2], Niranjana P.[2,3], and Rubens E. G. Machado[18]
for The LSST Star Clusters Working Group

[1] Departamento Acadêmico de Física, Universidade Tecnológica Federal do Paraná, Av. Sete de Setembro 3165, Curitiba, Brazil; nicholassouza@alunos.utfpr.edu.br
[2] Laboratório Interinstitucional de e-Astronomia (LIneA), Rua Gal. José Cristino 77, Rio de Janeiro, RJ 20921-400, Brazil
[3] Instituto de Física, Universidade Federal do Rio Grande do Sul, Av. Bento Gonçalves 9500, Porto Alegre, RS 90040-060, Brazil
[4] Centre for Astrophysics Research, University of Hertfordshire, College Lane, Hatfield, AL10 9AB, UK
[5] Department of Physics & Astronomy, University of North Carolina at Chapel Hill, NC 27599-3255, USA
[6] International Centre for Radio Astronomy Research, Curtin University, GPO Box U1987, Perth, WA 6845, Australia
[7] Facultad de Ciencias Astronómicas y Geofísicas de la Universidad Nacional de La Plata, and Instituto de Astrofísica de La Plata, Paseo del Bosque S/N, B1900FWA La Plata, Argentina
[8] Consejo Nacional de Investigaciones Científicas y Técnicas, Godoy Cruz 2290, C1425FQB, Ciudad Autónoma de Buenos Aires, Argentina
[9] INAF Osservatorio Astronomico d'Abruzzo, Via Maggini, 64100 Teramo, Italy
[10] Université de Strasbourg, CNRS, Observatoire astronomique de Strasbourg, UMR 7550, 67000 Strasbourg, France
[11] Department of Astronomy, Indiana University, Bloomington, IN 47405, USA
[12] Department of Physics & Astronomy, The University of Western Ontario, London, ON N6A 3K7, Canada
[13] IPAC, Caltech, 1200 E. California Boulevard, Pasadena, CA 91125, USA
[14] Waterloo Centre for Astrophysics, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada
[15] Perimeter Institute for Theoretical Physics, Waterloo, Ontario N2L 2Y5, Canada
[16] Instituto de Estudios Astrofísicos, Facultad de Ingeniería y Ciencias, Universidad Diego Portales, Av. Ejército Libertador 441, Santiago, Chile
[17] Centre national d'études spatiales (CNES), 2, Place Maurice Quentin, 75039, Paris, France
[18] Instituto de Astronomia, Geofísica e Ciências Atmosféricas, Universidade de São Paulo, Rua do Matão 1226, São Paulo, SP, Brazil
Received 2025 December 18; revised 2026 May 21; accepted 2026 May 24; published 2026 June 29

Abstract

Globular clusters (GCs) are excellent tracers of their host galaxies' evolutionary histories. Traditional methods for identifying GCs in galaxies rely on cuts over photometric catalogs and can yield source lists with high levels of contamination from compact background galaxies and foreground stars. In an era when large-scale sky surveys produce photometry for millions of sources, it is essential to employ flexible and scalable tools to reliably identify GCs in external galaxies. To prepare for surveys like Rubin/LSST, we need to explore practical methodological improvements and quantify the limitations inherent in datasets. This paper investigates the selection of point-like extragalactic GCs exclusively in the *ugrizY* color space. We use archival data to assemble an LSST-like photometric catalog for the Fornax Cluster containing labeled confirmed GCs, galaxies, and stars. From this catalog, using principal component analysis and nonlinear autoencoders (AEs), we construct inputs to random forest and multilayer perceptron classifiers. We show that selecting GCs using all 15 available colors can lead to a minimum contamination rate of ∼30%, whereas the use of color–color diagrams may double said rate. If only the first four principal components of the colors are used instead, the same minimum contamination rate is achieved without increasing incompleteness. The AEs did not improve GC identification. To further reduce contamination and extract the full potential of LSST for star-cluster studies, we argue for the need to augment photometric information with ancillary data (morphology from space-based missions and near-infrared photometry) before attempting to leverage more complex models.

Unified Astronomy Thesaurus concepts: Globular star clusters (656); Classification (1907); Dimensionality reduction (1943); Random Forests (1935); Neural networks (1933)

1. Introduction

A globular cluster (GC) is a very dense set of thousands to millions of stars, typically spanning masses from 10^4 to $10^6\,M_\odot$ (W. E. Harris et al. 2014; H. Baumgardt & M. Hilker 2018). GCs found in local galaxies are typically very old, with ages comparable to a Hubble time (C. Usher et al. 2019; K. Fahrion et al. 2020). They can be found in a wide variety of galaxy morphologies and carry rich information about the distant past of their hosts, thus allowing for the study of galaxy evolution (M. A. Beasley 2020). They have been observed to have half-light radii within the range $r_h \sim 2$–5 pc (K. L. Masters et al. 2010; J. P. Brodie et al. 2011; J. P. Caso et al. 2014), which means that for the vast majority of telescopes, GCs rapidly become point-like sources, the more distant the observed galaxy is.

The correct identification of extragalactic GCs, as well as their incorporation into simulations, allows for a wide variety of scientific applications. It is possible to place constraints on the amount and distribution of baryonic and dark matter from

the masses and spatial distributions of the GC systems of particular galaxies or galaxy clusters. When such results are interpreted in light of the properties of the host galaxies, pieces of the mass assembly history of these systems are revealed (W. E. Harris et al. 2013; M. J. Hudson et al. 2014; D. A. Forbes et al. 2018; A. Burkert & D. A. Forbes 2020; L. M. Valenzuela et al. 2021; T. Saifollahi et al. 2022; D. Zaritsky 2022; J. M. Diego et al. 2023; M. Reina-Campos et al. 2023; M. A. Canossa-Gosteinski et al. 2024; M. Mirabile et al. 2024; V. Dornan & W. E. Harris 2025). Also, distance estimates can be derived via the GC luminosity function (GCLF) if a universal form is assumed (M. Rejkuba 2012). In a complementary manner, kinematics and stellar population properties of extragalactic GCs are studied with spectroscopy and/or multiband photometry to further detail the evolutionary paths of their host galaxies (A. L. Chies-Santos et al. 2011a; F. Annibali et al. 2018; C. Usher et al. 2019; A. L. Chies-Santos et al. 2022; D. A. Forbes et al. 2022; A. Adamo et al. 2023; N. Grasser et al. 2024; L. Lomelí-Núñez et al. 2024; C. Usher et al. 2024). Beyond this, GCs are also known to host unique high-energy, transient phenomena, from ultraluminous X-ray sources (T. J. Maccarone et al. 2007; K. C. Dage et al. 2020) to fast radio bursts (F. Kirsten et al. 2022). As the study of GC systems has implications for a broad range of science cases, it is imperative to develop robust methodologies to identify extragalactic GC candidates.

Several systematic surveys that investigate extragalactic GC systems (among other objects) have been carried out in recent decades: e.g., the Hubble Space Telescope Advanced Camera for Surveys (HST ACS) Virgo Cluster Survey (P. Côté et al. 2004); the ACS Fornax Cluster Survey (ACSFCS; A. Jordán et al. 2007); the optical-near-infrared survey of GCs in early-type galaxies by A. L. Chies-Santos et al. (2011b); the Next Generation Virgo Cluster Survey (L. Ferrarese et al. 2012); the SAGES Legacy Unifying Globulars and GalaxieS (SLUGGS) Survey (J. P. Brodie et al. 2014); and the Fornax Deep Survey (FDS; M. Cantiello et al. 2020). Although different surveys develop distinct procedures to extract samples of extragalactic GC candidates, each with its specificities, there is a common ground, and some basic steps are well established. Since GCs can appear within the innermost parts of a galaxy out to many effective radii, a typical first step is to remove the diffuse stellar emission from the host galaxy. This is done either by fitting and subtracting a model of the galaxy light distribution or by smoothing the image and subtracting it from the original. A source-detection algorithm (e.g., the one available in `SExtractor`) is then used to find objects that have fluxes above a specified threshold compared to the background noise. Photometry is then performed, and a catalog of sources is produced, including their magnitudes and other quantities measured in the available filters.

The full catalog of the studied region contains, a priori, a heterogeneous population of unlabeled sources, including foreground stars, galaxies spanning a wide range of redshifts, and GCs. The methodology to identify the sources of different nature varies among different authors in the literature. The traditional approach is to apply linear cuts in the spaces of spectral energy distributions (SEDs), colors, and morphometric quantities. That is, lines and polygons are traced on color–color, color–magnitude diagrams (and other projections using, for example, FWHM, ellipticity) to define the regions associated with GCs and label candidates. This is the case for all the surveys previously cited, as well as other studies, including more recent works (J. R. Hargis & K. L. Rhode 2012; R. D'Abrusco et al. 2016; J. M. Berkheimer et al. 2025; S. Lim et al. 2025; T. Saifollahi et al. 2025a). On the other hand, methods not restricted to linear cuts have also been tested. Instead of tracing lines and polygons, it is possible to use curves that mix linear and nonlinear cuts to define boundaries. For instance, T. Saifollahi et al. (2025b) uses nonlinear cuts (obtained by comparisons with artificial GCs) to select candidates in the space of magnitudes, colors, compactness index, and ellipticity, and H. Dou et al. (2025) fits ellipses to define the GC locus in the $g - r$ versus $r - z$ diagram, performing linear cuts on the morphological parameters. Another alternative is to use a statistical technique such as propensity score matching (D. E. Ho et al. 2007; P. C. Austin 2011), so that sources are assigned as candidates based on the neighbors of each confirmed GC (A. L. Chies-Santos et al. 2022). Furthermore, machine learning models can be trained to identify GC candidates by exploring the relevant parameter space in a substantially less constrained manner than traditional approaches, based on linear cuts in a small number of observables. Examples include Gaussian mixture models (R. Garcia-Dias et al. 2020), random forest classifiers (L. Breiman 2001; G. Biau & E. Scornet 2015), and neural networks. These techniques have recently been applied to GC identification in several studies (E. Barbisan et al. 2022; D. Dold & K. Fahrion 2022; M. Mohammadi et al. 2022; Euclid Collaboration et al. 2025). Related applications to stellar population classification can also be found in the context of young stellar object identification (M. A. Kuhn et al. 2021).

Regardless of the method utilized to select extragalactic GC candidates, the final selection continues to contain an important percentage of contamination from mostly background galaxies and foreground stars, which display photometric and morphological signatures similar to those of GCs. Each of the approaches mentioned aims to reduce contamination by making a suitable comparison between the properties of GCs from the literature and the unlabeled data (sources of unknown nature) within a parameter space defined by the available data of each study. Spatial resolution and photometry depth are crucial to decreasing the number of contaminants, e.g., with HST ACS, as in A. Jordán et al. (2015). For the case of the Euclid Space Telescope, by incorporating morphometric information (FWHM, ellipticity, concentration index, etc.) in the GC selection pipeline, it is expected that contamination from background sources can be reduced by 90% (Euclid Collaboration et al. 2025). Another valuable type of information that can be used to reduce contamination is kinematics. Data from the Gaia mission (Gaia Collaboration et al. 2016), including parallax and proper motion measurements, can help discriminate extragalactic GCs from foreground stars (A. K. Hughes et al. 2021; A. L. Chies-Santos et al. 2022). In terms of SEDs and colors, it is impractical to unequivocally identify GCs from optical photometry in a few bands alone. High contamination can be significantly reduced by including near-ultraviolet and/or near-infrared information, using the u and/or K bands (A. L. Chies-Santos et al. 2011b; R. P. Muñoz et al. 2014; M. Cantiello et al. 2018; T. Saifollahi et al. 2021). Ultimately, the more bands considered, the more extensive the coverage of the electromagnetic spectrum and the more accurate the information on the nature of the objects. In this sense, the

definitive tool to confirm the nature of GC candidates is spectroscopy, which is, in turn, only available for a very limited number of systems, mostly nearby, with very bright GCs (J. P. Brodie et al. 2014).

The task of extragalactic GC candidate selection will face an unprecedented amount of high-quality data in the next few years, with the advent of next-generation multiband photometric surveys and telescopes such as the Vera C. Rubin Observatory's Legacy Survey of Space and Time (LSST; Ž. Ivezić et al. 2019), ESA's Euclid Space Telescope (Euclid Collaboration et al. 2022), NASA's Nancy Grace Roman Space Telescope (R. Akeson et al. 2019), and the Chinese Space Station Telescope (CSST Collaboration et al. 2025). The large volume of upcoming data requires us to adopt new approaches, in particular by looking to machine learning models. Our work occurs in the context of preparation for Rubin/LSST; we seek to contribute to the development of tools and predictions to be applied to future LSST data.

The Vera C. Rubin Observatory is a ground-based facility with an 8.4 m diameter (primary mirror) telescope, upon which the largest digital camera ever produced has been installed. The LSST camera comprises 3.2 gigapixels, with a pixel scale of $0\farcs2\,\mathrm{pixel}^{-1}$ and 9.6 square degrees of field of view. The purpose of LSST is to iteratively map the entire southern sky to achieve very deep photometry in the 6-filter *ugrizy*,[19] with an average cadence of 3 days over 10 yr of operation. A combination of the specifications of Rubin and the LSST strategy with the literature on the GCLF (with a peak at $M_g \sim -7.5$; A. Jordán et al. 2007) reveals that with a single estimated exposure, Rubin will be able to reach the GCLF turnover magnitude (TOM) in the *g* band of GC systems at distances up to 25 Mpc and 0.5 mag fainter than the TOM of systems up to ∼20 Mpc (about the distance to the Fornax Cluster). With coadded images over 10 yr of operation, again in the *g* band, it will reach the TOM of systems at 100 Mpc and 1.5 mag brighter than the TOM of systems at 150 Mpc. For the entire LSST footprint, in the expected final coadded images, C. Usher et al. (2023) estimates the detection of light from $\sim10^7$ GCs in the *griz* bands, $\sim10^6$ in the *y* band, and $\sim10^5$ in the *u* band. However, due to its limitations in terms of resolution, GCs beyond 10 Mpc are all expected to be point-like sources in Rubin images. Therefore, it is indispensable to seek methodological improvements for the correct identification of these remote point-like GCs that will be present in upcoming LSST data releases.

The ideal scenario would be to develop a scalable, adaptable model, capable of constructing clean samples of extragalactic GC candidates (resolved and unresolved) in and around a variety of galaxies, without the need for further galaxy-by-galaxy refinements, and calling external databases from other facilities (HST, Euclid, Roman, Gaia, etc.), which are only available for specific targets and restricted regions. However, publicly available data that could be used to train and test the reliability of such a model are highly heterogeneous, which leads to nonoptimal training and unreliable classifications: Data from different surveys, with very different characteristics and very few spectroscopically confirmed GCs, do not allow for robust and thorough experiments in supervised setups.

Given the limitations of optical photometry for selecting GCs and the fact that Rubin lacks both near-infrared filters and sufficient resolution to effectively distinguish GCs from other contaminants, the purpose of this paper is to answer the following question. With an LSST-like, *ugrizY* photometric catalog, relying exclusively on color information,[20] how well can point-like extragalactic GCs be distinguished from background galaxies and foreground stars (the main contaminants)? An equivalent way of framing our goal is, Do traditional color–color diagrams capture all of the clustering capability available in such a multiband photometric catalog? If not, i.e., assuming the color space in the catalog does contain sufficient information to further distinguish point-like GCs from contaminants (beyond what is achieved with 2D color–color diagrams), then there must exist a transformation capable of making this information evident, enhancing the contrast between the different classes of objects. In that context, we investigate whether this is the case or not.

Toward our goal, we assemble an LSST-like, photometric catalog containing labeled confirmed GCs from the literature, as well as background galaxies and foreground stars, as described in Section 2. Then, as detailed in Section 3, we transform the colors available in this dataset in two different ways to address the questions posited previously: (i) using principal component analysis (PCA; I. T. Jolliffe & J. Cadima 2016) and (ii) training nonlinear autoencoders (AEs; D. Bank et al. 2020). For our purposes, these techniques serve as dimensionality reduction tools (see, e.g., R. S. de Souza et al. 2022, 2014; Q. Xu et al. 2023). We then use machine learning classification models to test whether these transformations on the colors can improve the identification of the GCs in comparison with the use of the colors themselves as input to the models. The results of our tests are presented and discussed in Section 4. We summarize our results and suggestions in Section 5.

2. Data

Since LSST will provide data in the six-band *ugrizy*, we choose publicly available multiband photometry data of the Fornax Cluster provided by the Dark Energy Survey (DES) in the *grizY* bands (T. M. C. Abbott et al. 2021), alongside European Southern Observatory's Very Large Telescope (VLT) Survey Telescope (ESO VST) FDS, in the *ugri* bands (M. Cantiello et al. 2020), to compose an LSST-like photometric catalog.

2.1. Fornax Deep Survey Data and Confirmed Globular Clusters

The FDS is a deep imaging survey performed with the ground-based ESO VST, a 2.6 m diameter telescope at Cerro Paranal, Chile. It used the OmegaCAM camera to obtain images of 26 square degrees of the Fornax Cluster in the four bands *ugri*, with a pixel scale of $0\farcs21\,\mathrm{pixel}^{-1}$, and a field of view of 1.0 square degrees (R. Peletier et al. 2020). The FDS photometry described in M. Cantiello et al. (2020) uses a multiband coadded image created by stacking the best quality (sharpest FWHM) coadded single-band images in *gri*. `SExtractor` receives this stack image as input to derive properties such as the mean FWHM, `CLASS_STAR`, and flux

[19] The letter *y* is used exclusively to refer to the corresponding photometric filter present in the Rubin Observatory LSST Camera filter system. It is similar, but not identical, to the *Y* filter used in the Dark Energy Survey, which will be referenced frequently in this paper.

[20] As a starting point, we aim to understand the specific role of Rubin/LSST colors for the identification of GCs. Thus, for the time being, we abstract away from SEDs and morphometric measurements, considering them as additional information to eventually complement the color space.

Table 1
FDS Data Quality Information from M. Cantiello et al. (2020)

	u	*g*	*r*	*i*
Magnitude limit (mag)	24.1	25.4	24.9	24.0
PSF FWHM (arcsecond)	1.26	1.12	0.92	0.94

Note. The first row reports limiting magnitudes derived from 5σ magnitude integration over the PSF. The second row presents the median FWHM.

radius. Moreover, `DAOphot` is used in this image to model the point-spread function (PSF) and identify sources. Magnitudes are then estimated by integration over the PSF in each single-band coadded image. We access the resulting catalogs through the VizieR catalog access tool (F. Ochsenbein et al. 2000). All the magnitudes in question are in the AB system. Important quantities about FDS observations and photometry are presented in Table 1.

Furthermore, M. Cantiello et al. (2020) matched the FDS *ugri* photometric catalog with the spectroscopic samples of GCs in Fornax produced by Y. Schuberth et al. (2010) and the Fornax Cluster VLT Spectroscopic Survey (V. Pota et al. 2018). In this way, 1342 spectroscopically confirmed GCs, the majority associated with NGC 1399, are labeled in the FDS catalog. Additionally, another 1921 sources are labeled as photometrically confirmed GCs due to their marginally resolved appearances in ACSFCS images (A. Jordán et al. 2007), located in other galaxies across the Fornax Cluster. As described in M. Cantiello et al. (2020), these 1921 sources are selected from a cut on the probability of being a GC $p_{\rm GC} > 0.75$; see A. Jordán et al. (2009) for a detailed description of this probability measure. Still, of these photometrically confirmed GCs, 214 have $p_{\rm GC}$ below 95%, 110 are below 90%, and only 23 are below 80%. Finally, A. Chaturvedi et al. (2022) increased the number of spectroscopically confirmed GCs in the Fornax Cluster central region. Out of their catalog of confirmed GCs, 296 were previously unidentified in the catalog of all FDS sources from M. Cantiello et al. (2020); hence we use FDS IDs to label them accordingly. In this work, we take into consideration all of the 1342 + 268 spectroscopically confirmed GCs, together with the 1921 photometrically confirmed ones, hereafter referred to as simply "confirmed GCs."

As the scope of this paper is to compare the effects of different methods of transforming colors to identify GCs, rather than to select new GC candidates, we restrict our analysis to the sources found within the circular region with a radius of 1° around NGC 1399, within which the photometric calibration across bands (and therefore colors) is the most homogeneous. This region contains all of the spectroscopically confirmed GCs mentioned previously and a few hundred of the photometrically confirmed ones. See Figure 1 for a visualization of the spatial distribution of the sources. Given the instrumental details of FDS and the distance to the Fornax Cluster (∼19.3 Mpc; G. S. Anand et al. 2024), GCs are point-like sources (M. Cantiello et al. 2020).

2.2. Dark Energy Survey Data and the Selection of Stars and Galaxies

The DES is a ground-based, wide-area visible, and near-infrared imaging survey that covers approximately 5000 square degrees of the southern sky in the *grizY* bands. DES imaging is performed with the Dark Energy Camera (DECam), which is mounted on the 4 m Blanco Telescope at Cerro Tololo Inter-American Observatory in Chile (T. M. C. Abbott et al. 2021), with a pixel scale of $0.''264$ pixel^{-1} and a field of view of 3 square degrees. We use the DES Data Release 2 (DR2) photometric catalog obtained via `SExtractor` (E. Bertin & S. Arnouts 1996) in double-image mode, where the detection image is the linear combination stack of coadded images in the three bands $r + i + z$. DES DR2 contains magnitude estimations that use several aperture models. The most important for this work are `MAG_AUTO` (elliptical model based on the Kron radius) and `MAG_APER` (circular apertures). We disregard DES weighted-average PSF photometry upon realizing that, for the 1 degree radius circular region around NGC 1399, it contains large fractions of missing values across all *grizY* bands—respectively, 52%, 43%, 51%, 63%, and 80%. Such high fractions are a consequence of the fact that weighted-average PSF magnitudes are measured only for sufficiently bright sources that are detectable in single-epoch images (T. M. C. Abbott et al. 2021). All DES magnitude values are in the AB system.

We access DES data through the Astro Data Lab science platform (R. Nikutta et al. 2020), and as explained in Section 2.1, we select sources that lie within the circular area of a 1° radius around NGC 1399, which are shown in Figure 1—a total of 395,813 sources. Some relevant quantities on the DES DR2 data are presented in Table 2. As they are in the FDS images, GCs in the Fornax Cluster are point-like sources in the DES DR2 images.

Beyond labeling spectroscopically confirmed GCs and joining DES and FDS photometric catalogs, we also have to label samples of background galaxies and foreground stars. The selection criteria must ensure minimum contamination from unidentified GCs, which could otherwise compromise the interpretation of the performance of the classification models later on. Toward this end, we leverage the galaxy–star separation criteria suggested by T. M. C. Abbott et al. (2021) to select clean samples of galaxies and stars from DES DR2 data. These criteria combine a morphology-based classification variable named `EXTENDED_COADD`, which in turn is based on `SExtractor` `SPREAD_MODEL`, with the magnitude values `mag_auto_i`. We use altered versions of the selection criteria presented by T. M. C. Abbott et al. (2021), now considering fixed values of `EXTENDED_COADD` in each selection and modifying the `mag_auto_i` cut from [19.0, 22.5] to [18.0, 20.0] for the stellar selection:

$$\text{Galaxy selection: } \texttt{EXTENDED\_COADD} = 3 \ \& \ 19.0 \leqslant \texttt{mag\_auto\_i} \leqslant 22.5 \quad (1)$$

$$\text{Stellar selection: } \texttt{EXTENDED\_COADD} = 0 \ \& \ 18.0 \leqslant \texttt{mag\_auto\_i} \leqslant 20.0 \quad (2)$$

The values of 3 and 0 for `EXTENDED_COADD` correspond to high-confidence galaxies and stars, respectively, whereas the original criteria consider the ranges $\texttt{EXTENDED\_COADD} \geqslant 2$ and $0 \geqslant \texttt{EXTENDED\_COADD} \geqslant 1$ to include the "likely" galaxies and stars in the selections. We refer to Section 4.5 of T. M. C. Abbott et al. (2018) for a detailed description of the connection between `EXTENDED_COADD` and `SPREAD_MODEL`, and for the reason why such variables may render more reliable morphological classifications compared to those derived using `CLASS_STAR`.

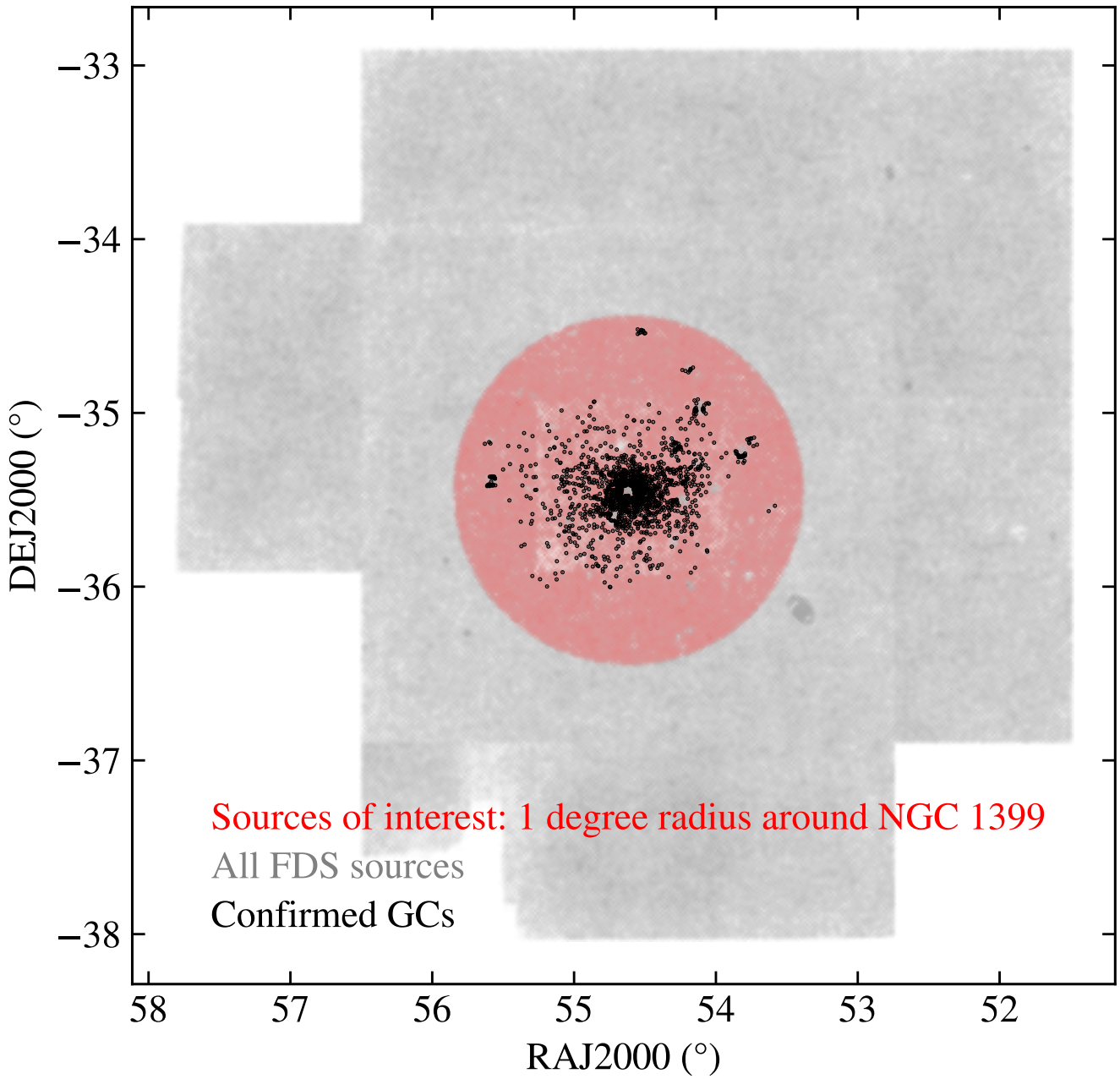

Figure 1. Entire coverage of FDS in gray; the red circle indicates the 1° radius region around NGC 1399, our sources of interest. The black points represent the positions of confirmed GCs for which we have FDS *ugri* and DES *grizY* photometry available.

Table 2
DES DR2 Data Quality Information from T. M. C. Abbott et al. (2021)

	g	*r*	*i*	*z*	*Y*
Magnitude limit (mag)	24.7	24.4	23.8	23.1	21.7
PSF FWHM (arcseconds)	1.11	0.95	0.88	0.83	0.90

Note. The first row displays the median limiting magnitudes of the coadded catalog for a 1.″95 diameter aperture (`MAG_APER_4`), at S/N = 10.

We now discuss the motivation behind the stellar selection criteria modification, as well as the possible contamination from unlabeled GCs in our galaxy and star samples. As the GCs in question are all point-like sources, `EXTENDED_COADD` must provide enough information to select a very pure sample of extended background galaxies; no contamination from point-like sources is expected in this case. Regarding our stellar sample, the original selection criterion from T. M. C. Abbott et al. (2021) considers $19.0 \leqslant$ `mag_auto_i` $\leqslant 22.5$, which arguably leads to an appreciable portion of point-like GCs in Fornax being selected as stars. Namely, at the distance of Fornax, $m - M \sim 31.5$, and assuming an absolute GCLF TOM in the i-band $\mathrm{TOM}_i \sim -8.5$ mag, we obtain an apparent $\mathrm{aTOM}_i \sim 23$ mag. Now, with a GCLF spread $\sigma_{\mathrm{GCLF}} \sim 1.5$ mag for the entire Fornax cluster (which accounts for the cluster depth), we ascertain that the cut `mag_auto_i` $\leqslant 22.5$ is located $\sim 0.33\sigma_{\mathrm{GCLF}}$ from aTOM_i. This implies that the stellar sample would still contain roughly 35–40% of the entire GC population in Fornax. This fraction is reduced to less than 3% if the cut is at 20 mag. This is why we decided to modify this cut, as a way to

Table 3
The Expected 5σ Depths for LSST Single Exposure and Coadded Images (after the 10 Yr Survey), Estimated from Operation Simulations (F. B. Bianco et al. 2022)

	u	*g*	*r*	*i*	*z*	*y*
Single image lim. magnitude (mag)	23.8	24.5	24.0	23.4	22.7	22.9
Coadded image lim. magnitude (mag)	25.6	26.9	26.9	26.4	25.6	24.8

ensure that unidentified GCs represent a very small, negligible fraction of the total sample of stellar objects. We also shift the bright end of the cut from 19.0 to 18.0 so that the selection is not too restrictive. The actual selection of galaxies and stars is performed with an extra ad hoc constraint: Exclude all sources currently labeled as confirmed GCs. The final number of selected galaxies and stars is discussed in Section 2.4.

T. M. C. Abbott et al. (2021) also states that following Criterion (1) and the original version of Criterion (2), galaxies and stars can be selected with efficiency rates greater than 99% and 94%, and contamination rates lower than 2% and 3%, respectively. We stress that such values are derived in relation to the entirety of HSC SSP Data Release 1 (DR1; H. Aihara et al. 2018) and thus do not imply that the use of `SPREAD_MODEL` is sufficient to break nearly all degeneracies in the identification of compact objects such as GCs, as discussed previously.

After crossmatching FDS and DES catalogs, we apply Criteria (1) and (2) to label extended background galaxies and point-like foreground stars in the resulting matched catalog, as described in Sections 2.3 and 2.4.

2.3. LSST-like Dataset: Combining DES and FDS Data

As described in Section 1, LSST will measure the light of an enormous number of extragalactic GCs; it will detect GCs out to ∼200 Mpc (C. Usher et al. 2023). The expected LSST limiting magnitudes for single images and coadded ones (after 10 yr of observations) are shown in Table 3.

Comparing Tables 1, 2, and 3, we see that FDS achieves magnitude limits deeper than DES, while both are deeper than the expected limiting magnitudes of single exposures of LSST, except in the *y* band. However, the expected photometric depths of coadded images after 10 yr of the LSST survey are well beyond those of FDS and DES. Moreover, the values of the pixel scales and median FWHMs of sources in DES and FDS are very similar, suggesting that the combination of their photometric data should not create compromising biases in terms of data quality. The details of this combination are now discussed.

First, we perform a crossmatch with the sky coordinates in the DES and FDS catalogs. An angular separation tolerance of 1.″0 was used based on the median FWHM values of the surveys. Of the 1342 spectroscopically confirmed GCs present in the catalog of M. Cantiello et al. (2020), 1129 are also available in DES DR2 catalog. Of the 292 spectroscopically confirmed GCs from A. Chaturvedi et al. (2022), after filtering out those whose spectra have $S/N < 3$, 268 are also present in the DES DR2 catalog. Finally, 245 photometrically confirmed GCs are in the matched catalog, giving us a total of 1642 confirmed GCs at this stage. The missing GCs in question are all located within the very bright central region of their host galaxies (including NGC 1399), which are disregarded by the source extraction and deblending routines of the DES pipeline (T. M. C. Abbott et al. 2018). It is important to emphasize that the strategy used to extract and measure the sources in the FDS catalog was designed toward compact systems science, with the goal of studying GCs and ultracompact dwarf galaxies. Meanwhile, DES aims to produce consistent measurements over a large number of observations made throughout ∼5000 square degrees of the southern sky. This is consistent with the number density of sources in the Fornax Cluster being greater in FDS than in DES. All mentions of "catalog data" hereafter refer to the matched catalog.

Although only FDS has data in the *u* band, and the *z* and *Y* bands are available only in DES, the two surveys share the *gri* bands. Beyond that, DES and FDS provide magnitudes using various aperture models. To decide which magnitude values to use for our analysis, it is necessary to further compare the photometry from FDS and DES as follows. All magnitude values in both FDS and DES were corrected by reddening in accordance with D. J. Schlegel et al. (1998) and E. F. Schlafly & D. P. Finkbeiner (2011).

We first consider only the *gri* bands, available in both surveys. In Figure 2, the thinnest exponential scatters indicate that DES circular aperture photometry has the lowest errors among the other models for these three bands. The spectroscopically confirmed GCs also have the lowest magnitude errors in the circular aperture model. `MAG_APER_5` corresponds to 11.11 pixels or 2.″92 of diameter, but we also studied the error curve for smaller circular apertures of 1.″95 and 1.″46 (`MAG_APER_4` and `MAG_APER_3`, respectively). They are qualitatively indistinguishable from those presented in the middle row of Figure 2 and thus are not included in the figure. Furthermore, `SExtractor` magnitude error estimation is found to be more reliable compared with that of `DAOphot` (A. L. Chies-Santos et al. 2011b). Therefore, from the perspective of the magnitude error curves, the circular aperture model is favored over the automatic one. Still, there are many options for aperture size.

To choose the most suitable aperture size, we examine Figure 3, which shows the differences in magnitude values for the same sources, in the same band, but in different surveys and for different aperture models and sizes. We find that DES `MAG_APER_5` and `MAG_AUTO` are the models whose magnitudes deviate the least from the PSF photometry of FDS (taken here for reference), especially when considering sources brighter than 20 mag in the three bands. However, for fainter sources, DES `MAG_AUTO` values significantly differ from the FDS magnitudes compared to the circular aperture models, which are more consistent. For DES `MAG_APER_4` magnitudes, a vertical shift is observed: With this circular aperture, DES magnitudes are systematically fainter than those from the FDS. For DES `MAG_APER_5`, this shift is not as pronounced—that is, a larger aperture accounts for light that was not captured with `MAG_APER_4` but was picked by the PSF photometry of FDS. Statistics on these magnitude differences are presented in Table 4. Among the DES aperture models studied compared to FDS PSF photometry, the lowest means, medians, and root mean squares are found to be those associated with DES `MAG_APER_5`. We performed the same examination also considering smaller and larger circular apertures (e.g., `MAG_APER_3`∼1.″46 and `MAG_APER_6`∼3.″90), and the results are as expected: Smaller apertures disregard even more light

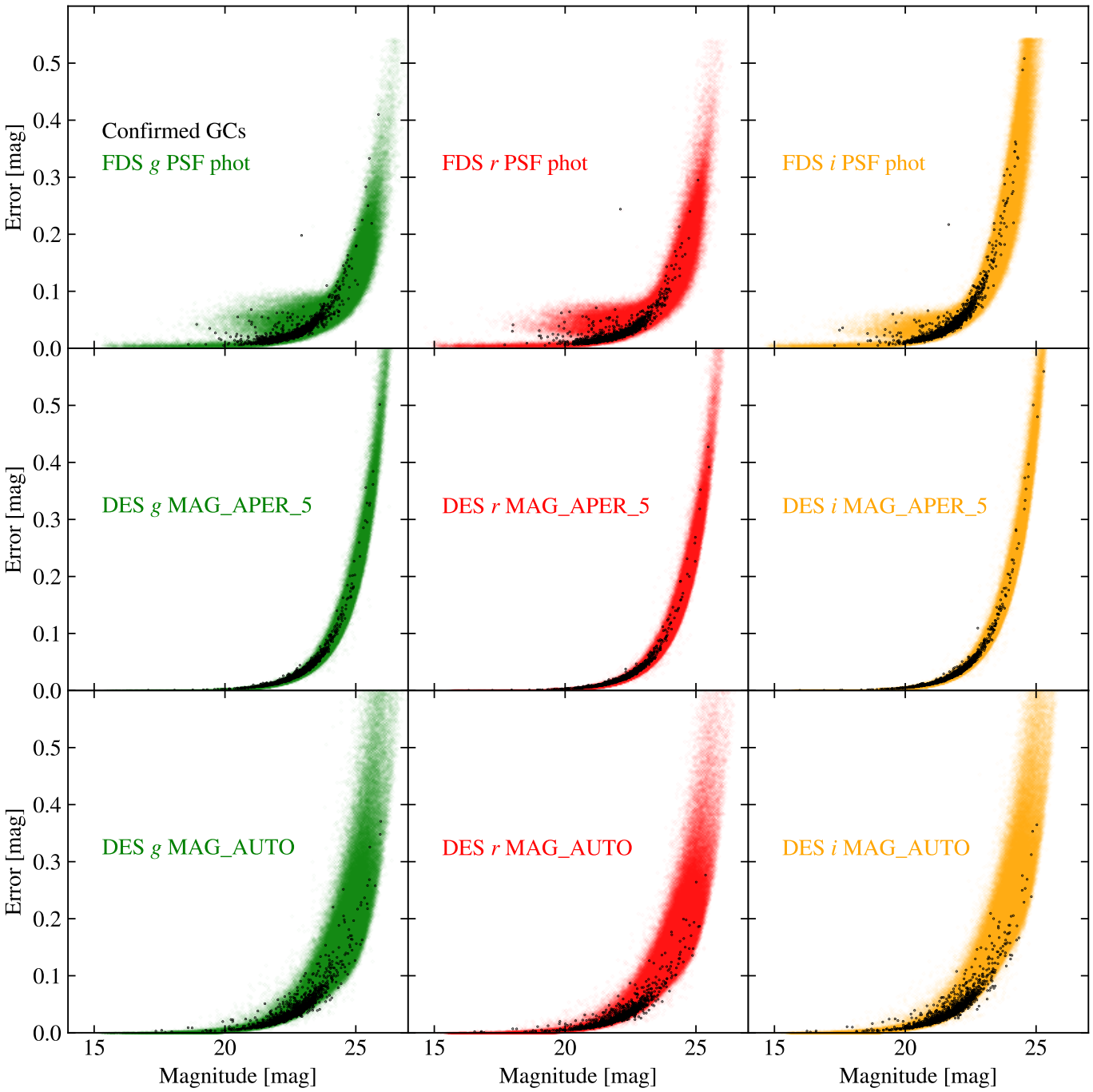

Figure 2. Magnitude errors plotted vs. magnitudes for the bands in common between DES and FDS, *gri*. Black dots represent confirmed GCs. The first row of plots refers to FDS PSF photometry data, the second to DES circular aperture photometry, and the third to DES automatic aperture (based on the Kron radius) photometry. For visualization purposes, the magnitude error axes were truncated at a value of 0.6 mag. Errors in FDS data do not exceed 0.6 mag: All FDS data points are visible in these plots.

compared with the case of `MAG_APER_4`, yielding even greater magnitude difference means, medians, and root mean squares. Meanwhile, apertures larger than that of `MAG_APER_5` also cause these statistical measures to increase in value, likely because apertures that are too large increase noise and/or capture part of the light of neighboring sources. Finally, the distributions of spectroscopically confirmed GCs for the DES `MAG_APER_{5,4}` cases in Figure 3 are more concentrated around the $y = 0$ line when compared with the `MAG_AUTO` case, once again favoring circular aperture over the automatic one.

Based on the provided investigation, we decided to compose our LSST-like *ugrizY* photometric catalog with the *u* band from FDS PSF photometry alongside DES *grizY* `MAG_APER_5` circular aperture photometry (11.11 pixels or $2\farcs92$ of diameter). No aperture correction was applied.

We are interested not only in confirmed GCs, but also in labeling background galaxies and foreground stars using Criteria (1) and (2). Thus we include the same plots as in Figure 3 in Appendix A, but highlighting the two categories of contaminants in Appendix A's Figures 8 and 9.

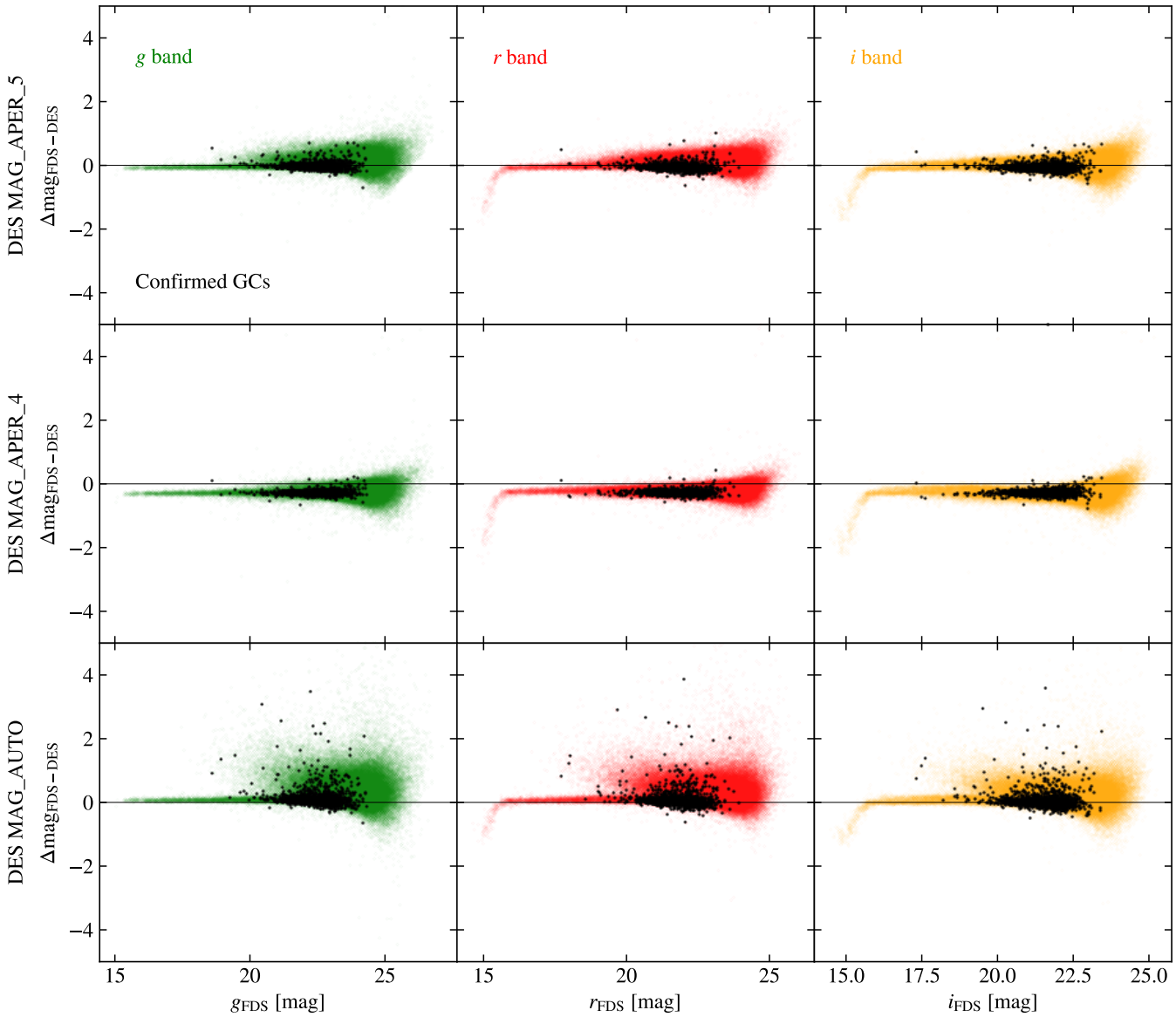

Figure 3. $\Delta\mathrm{mag}_{\mathrm{FDS-DES}}$ vs. g, r, i_{FDS}: the difference in magnitude for the same source, in the same band, but in different surveys compared to the FDS magnitude in the same band. The first row displays the plots where DES `MAG_APER_5` (2.″92) data were used, the second displays plots from DES `MAG_APER_4` (1.″92) data, and the third displays plots from the DES `MAG_AUTO` data. The horizontal black line is $y = 0$.

Table 4
Statistics on the Differences in Magnitude Values for the Bands in Common between FDS and DES

Filter	Mean $\Delta\mathrm{mag}_{\mathrm{FDS-DES}}$ (mag)	Median $\Delta\mathrm{mag}_{\mathrm{FDS-DES}}$ (mag)	rms $\Delta\mathrm{mag}_{\mathrm{FDS-DES}}$ (mag)	σ_{FDS}
g	(0.023, −0.318, 0.184)	(0.030, −0.308, 0.141)	(0.32, 0.40, 0.54)	1.26
r	(0.023, −0.272, 0.182)	(0.037, −0.264, 0.147)	(0.28, 0.38, 0.48)	1.40
i	(−0.042, −0.370, 0.052)	(−0.043, −0.340, 0.074)	(0.40, 0.48, 0.51)	1.44

Note. The values are displayed in triplets representing the different aperture models for DES photometry, while FDS PSF magnitudes are the same for each comparison (`MAG_APER_5`, `MAG_APER_4`, `MAG_AUTO`). σ_{FDS} is the standard deviation in FDS magnitudes, for reference.

2.4. Data Preprocessing: Filtering and Labeling Sources

Another important aspect to consider is the presence of missing values. In our catalog, about 9% of the data points miss the Y-band magnitude and less than 1% miss magnitude values in other bands. The FDS u band does not have any missing values. We removed any sources that lack a magnitude value in at least one band, leaving us with 190,281 sources, of which 1595 are confirmed GCs.

We filter out all sources that have at least one of the *ugrizY* bands with a magnitude error greater than 0.5 mag. This procedure leaves us with 105,318 sources in total. Lastly, we add labels to all remaining sources that fall within the DES selection criteria (Criteria [1] and [2]). These sources, together

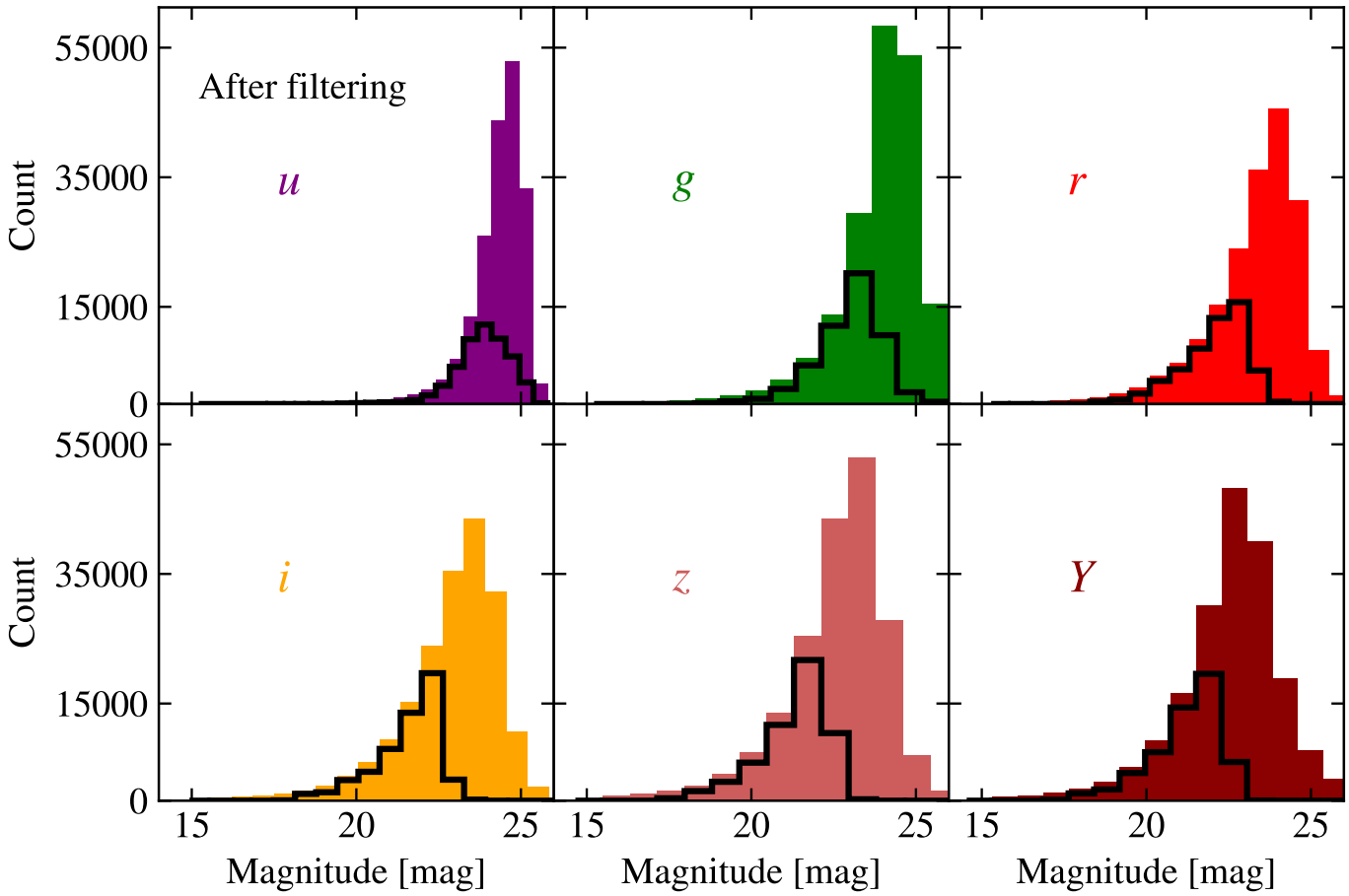

Figure 4. Distribution of magnitude values for each band. The colored bars represent the dataset before the preprocessing described in Section 2.4. The black edges indicate the subset that represents the dataset after all the filtering; it contains only the labeled sources.

with the confirmed GCs, comprise our final, LSST-like, filtered, and labeled dataset with three classes: 1440 confirmed GCs, 49,579 background galaxies, and 3726 foreground stars—54,745 sources in total. The fact that the different classes have very different numbers of associated data points (class imbalance) is taken into account when training the classification models. Figure 4 allows us to visualize the scale of the filtering we applied to the data.

We recall the discussion on our efforts to minimize the contamination from unidentified GCs in our samples galaxies and stars, as presented in Section 2.2. Background galaxies indeed dominate the total number of sources in our curated catalog, although no point-like GCs nor stars are expected to be present in this sample due to the role of `EXTENDED_COADD` in Criterion (1). The magnitude cut in Criterion (2), together with our knowledge of the GCLF of the Fornax Cluster, led us to expect unknown GCs to represent less than 3% of our stellar sample. In addition, we ran the TRIdimensional modeL of thE GALaxy (TRILEGAL) population synthesis code, introduced and calibrated by L. Girardi et al. (2005), using its default parameter values to obtain an estimate for the number of galactic stars within the 1° circular region around NGC 1399 from synthetic apparent magnitudes in the DES filters. After imposing the DES magnitude limits to filter our TRILEGAL output catalog and applying the same magnitude cut as in Criterion (2), we arrive at an estimated total of 4756 galactic stars within the sky region in question, which differs from the number of foreground stars in our sample by about 1000 counts. We interpret this difference to be barely within the expected uncertainty in star counts predicted by TRILEGAL, of about 20% (P. Dal Tio et al. 2022), placing our foreground star sample in borderline agreement with the model regarding star counts.

3. Methodology

Given the filtered and labeled dataset, we formulate the task as a classification problem aimed at assessing whether color transformations via dimensionality reduction enhance the separation of the confirmed GCs from the select background galaxies and foreground stars. We evaluate this using two classifiers: an RFC (L. Breiman 2001) and a multilayer perceptron classifier (MLPC; F. Murtagh 1991).

We consider both RFCs and MLPCs to probe the extent to which the information content of the LSST-like catalog requires nonlinear decision boundaries. RFCs provide a strong, well-calibrated baseline for tabular data, capturing feature interactions with minimal tuning and offering robustness to noisy photometry and heterogeneous source populations. MLPCs, by contrast, impose fewer structural assumptions and can represent smoother, higher-capacity nonlinear mappings in color space. Comparing their performances, therefore, serves as a diagnostic: A consistent gain from MLPCs would suggest that the class separation benefits from higher capacity, continuous nonlinear representations, whereas comparable performance would indicate that the discriminative signal is largely captured by lower-capacity tree ensembles (or is limited by measurement noise, selection effects, and label uncertainty), rather than by model expressiveness.

These models are trained and tested using three types of input: (i) colors derived from *ugrizY* photometry, (ii) the principal components (PCs) of these colors, and (iii) the latent space coordinates (LSCs) of AEs (D. Bank et al. 2020) that process the colors as input. For our purposes, AEs can be understood as neural networks designed to compress data efficiently—a dimensionality reduction tool, not limited to linear transformations (Q. Fournier & D. Aloise 2021), which is the case for PCA. For a more detailed description of AEs, see Section 3.1.

The core idea behind using PCA and AEs to transform the colors is that both techniques can potentially compact information into lower-dimensional representations, revealing and emphasizing clustering patterns in the data. This idea was also explored by, e.g., R. D'Abrusco et al. (2016) and A. L. Chies-Santos et al. (2022). R. D'Abrusco et al. (2016) selected GC candidates around NGC 1399 using PCA on a 3D color space of FDS catalogs, and reported its GC system to be extended so that it connects with those of neighboring galaxies. A. L. Chies-Santos et al. (2022) selected GC candidates in the M81/M82/NGC 3077 triplet using the PC1–PC2 diagram derived from the 12 dimensional SED space of J-PLUS photometric catalogs. Although no quantitative comparison with other methods was presented in these studies, they obtained solid lists of GC candidates, with their spatial distributions being interpretable in light of cluster-galaxy and galaxy–galaxy interactions. We aim to provide such a quantitative comparison.

Our choice to exclusively consider the color space instead of the combined space of SEDs and colors is justified as follows. During our preliminary tests, we noted that the use of colors as the only input (as opposed to using colors and SEDs) yields a slightly more compact distribution of GCs in principal component diagrams. In the context of our specific investigation, we interpreted this as an indication that the SEDs in question (each composed of only six data points) do not carry information capable of improving the clustering of GCs already observed in the color space or in the space of PCs of colors. Furthermore, associated with the previous discussion is the fact that our samples of labeled objects (GCs, background galaxies, and foreground stars) are biased toward brighter sources. Therefore, choosing inputs to be distance-independent quantities (such as colors) also serves as a way to mitigate this selection bias.

To evaluate the performance of the models, we use the output metrics precision (Equation [3]), recall (Equation [4]), and F1-score (Equation [5]). In these expressions, TP stands for true positives, FP refers to false positives, and FN refers to false negatives. Figure 5 synthesizes our analysis procedure.

$$\mathrm{P} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}) \tag{3}$$

$$\mathrm{R} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \tag{4}$$

$$\mathrm{F1} = 2\mathrm{PR}/(\mathrm{P} + \mathrm{R}) \tag{5}$$

3.1. Preparing Model Inputs

We aim to investigate whether more concise inputs could affect the identification of GCs. For this, we prepare input sets to the models using different numbers of colors, PCs, and LSCs.[21]

[21] This is equivalent to stating that we are interested in assessing the effects of dimensionality reduction on our classification problem.

3.1.1. Colors

The available colors are obtained by subtracting two magnitude values in different bands in all possible ways. We are working with the *ugrizY* filter set, which comprises 15 colors: $u-g$, $u-r$, $u-i$, $u-z$, $u-Y$, $g-r$, $g-i$, $g-z$, $g-Y$, $r-i$, $r-z$, $r-Y$, $i-z$, $i-Y$, $z-Y$. We define three input sets to both RFC and MLPC, composed of colors only. The first one comprises all 15 available colors as input. Among other outputs, RFCs provide a feature "importance" measurement, which uses internal metrics such as Gini importance or mean decrease in accuracy to quantify how much each input feature contributes to the prediction accuracy of the model. Using the feature importance values obtained when running the RFC with 15 colors, we choose the four "most important" colors, which were $u-g$, $u-z$, $u-i$, and $g-r$, to constitute our second color input set, a 4D input to both RFC and MLPC. Finally, using the same RFC feature importance ranking, the two most important colors, $u-g$ and $u-z$, are the third color input set.

3.1.2. Principal Components

The PCs of the colors are obtained via the Principal Component Analysis (PCA) class implemented in the Scikit Learn Python package, which in turn uses singular value decomposition to compute the PCs (F. Pedregosa et al. 2011). To suitably compare the model runs using colors and PCs as input to RFC and MLPC, we again consider three input sets: all 15 PCs, the first 4 PCs, and the first 2 PCs.

3.1.3. Latent Space Coordinates

We now discuss AEs and their latent spaces (D. Bank et al. 2020). The purpose of an AE, which is a neural network, is to compress, i.e., to encode the input data into a lower-dimensional representation (the latent space, represented by the innermost layer of the network), such that it contains enough information to reconstruct the original data up to a user-defined acceptable loss. In this sense, an AE is a dimensionality reduction tool (Q. Fournier & D. Aloise 2021). The part of the network responsible for the compression is called the *encoder*, while the one used to reconstruct the original data is named the *decoder*. Naturally, the target data used to compute the loss function and train the network must be identical to the input data: One desires to minimize the difference between the input data and its reconstructed version from the lower-dimensional latent space to obtain the best compression possible. The dimension of the latent space—that is, the dimension of the compressed version of the data—is to be defined by the user. Unlike the case of PCA, which outputs 15 linear combinations if 15 colors are given, for an AE, a latent space with as many dimensions as the input data is meaningless because no compression is achieved.

We decided to use two AEs, one with a latent space of four dimensions and the other with two, thus transforming the set of 15 colors into 4D and 2D representations. Again, this choice allows for more direct comparisons with the other 4D and 2D inputs composed of colors and PCs. Therefore, we make two runs with the coordinates of the colors in the latent spaces (LSCs) as input to each classification model.

Figure 6 shows our dataset projected into the three different spaces of interest: a usual color–color diagram, the PC1–PC2 diagram, and the 2D latent space of one of our nonlinear AEs (LS1–LS2 diagram). The plots show no qualitative differences

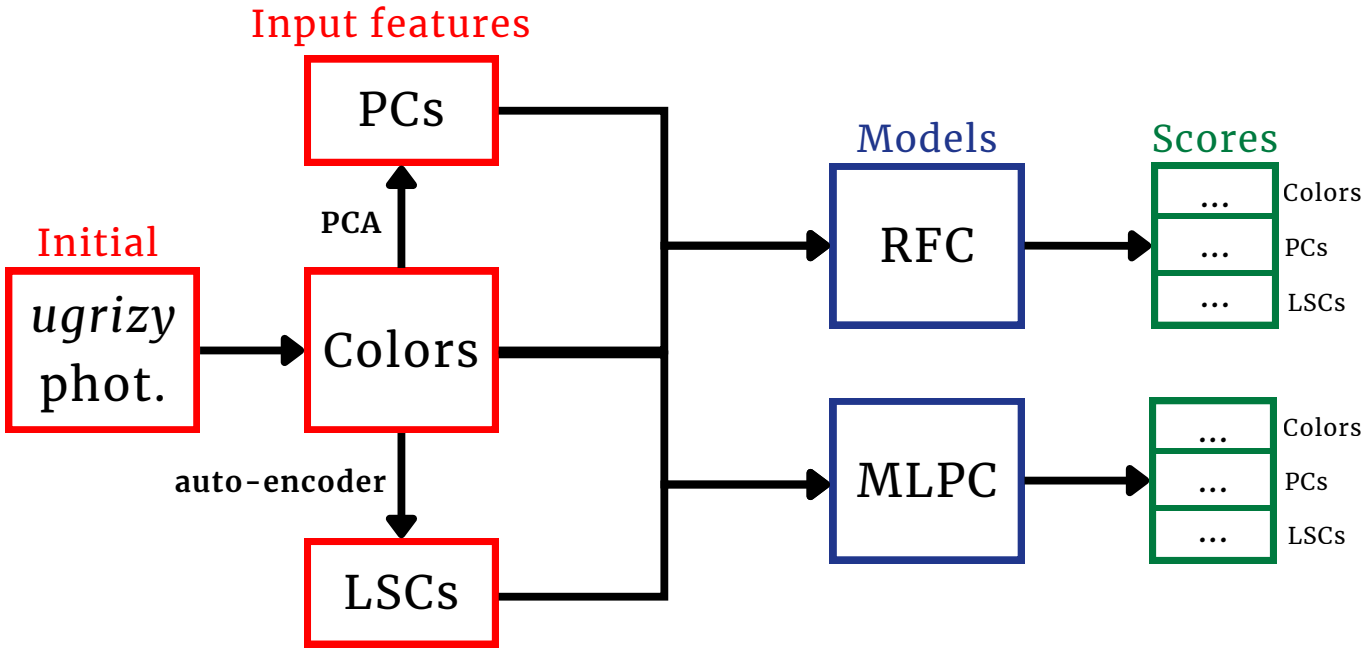

Figure 5. Diagram to illustrate the flux of data in our analysis procedure.

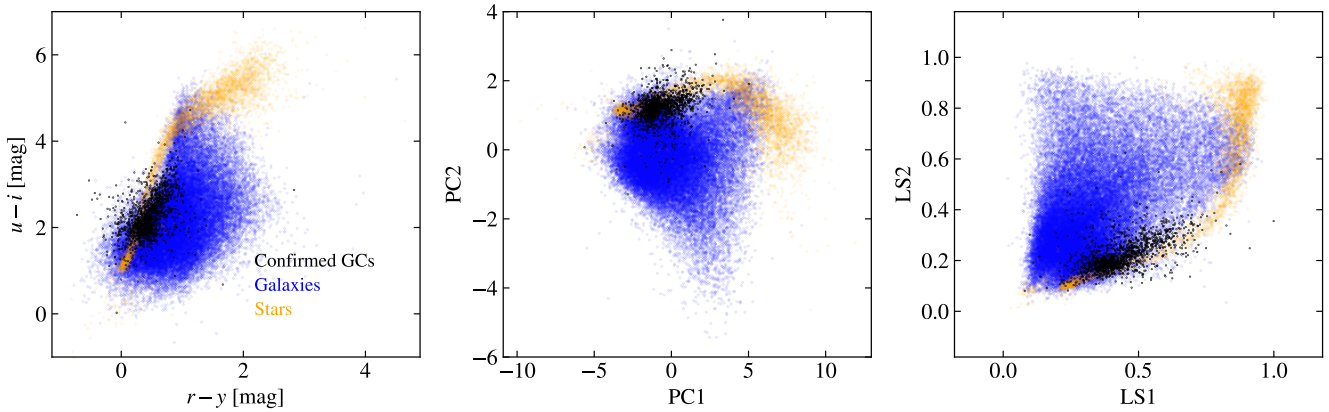

Figure 6. Projections in the color space, PC space, and the nonlinear AE latent space of our LSST-like filtered and labeled photometric catalog. The plots show that there are no qualitative differences in the distribution of the points in these spaces.

in the distribution of points in these spaces, which leads us to suspect in advance that transforming the colors will not facilitate the correct identification of GCs in this dataset. The role of the classification models is to allow for a quantitative evaluation of this suspicion.

3.2. Architecture Choices and Model Training

The specific architecture of the neural networks in question (AEs and MLPCs) depends on the dimension of the input data. Regarding the architecture of the two AEs, their input layers have 15 neurons, associated with the 15 input colors. The 15 dimensions of input are to be compressed into 4 in one of the AEs and 2 in the other. Hence the layers that will represent the latent spaces have 4 and 2 neurons, respectively. The output layers in both networks have, again, 15 neurons, as the AEs are designed to attempt to reconstruct the input. The encoders and decoders are chosen to be composed of only 1 encoding and 1 decoding layer, with 7 neurons each for the case of the 4D latent space and 5 neurons each for the 2D one. We choose the sigmoid function (nonlinear) as the activation of the encoder and a linear activation function for the decoder. The mean squared error loss is used to train the two AEs.

With respect to the MLPCs, their input layers have 15, 4, or 2 neurons (the 15 colors, 15 PCs or fewer, 4 or 2 LSCs), and the output layers have 3 neurons (as the problem has 3 classes). We decided to use 2 hidden layers, with 10 and 5 neurons, for the cases with 15 input features, 1 hidden layer with 8 neurons for the case of 4 input features, and 1 hidden layer with 4 neurons for the case of 2 input features. For all layers, except the output one, we use the ReLU activation function. For the output layer, we use the softmax function. The sparse categorical cross-entropy loss is used to train the MLPCs.

To properly handle class imbalance, all RFCs and MLPCs were trained using fivefold stratified cross-validation (CV) resampling (a technique used to train and evaluate models several times, each using a separate part of the training set), while maintaining an adequate proportion of class labels as the original dataset. The RFCs were also subjected to hyperparameter tuning via randomized search, allowing the number of trees to vary, as well as the maximum depth of the tree, the

Table 5
Results for the Classification of GCs: Performance Metric Triplets (Precision, Recall, F1-score) of the Best-performing Models (the Ones with Hyperparameter Values That Maximized the F1-score for the GC Class)

	RFC	MLPC
15 colors	(0.69, 0.29, 0.41)	(0.21, 0.88, 0.34)
4 colors	(0.55, 0.27, 0.36)	(0.20, 0.92, 0.33)
2 colors	(0.39, 0.22, 0.28)	(0.18, 0.91, 0.30)
15 PCs	(0.71, 0.29, 0.42)	(0.21, 0.91, 0.35)
4 PCs	(0.69, 0.29, 0.41)	(0.18, 0.90, 0.30)
2 PCs	(0.35, 0.17, 0.23)	(0.17, 0.86, 0.28)
4 LSCs	(0.66, 0.32, 0.43)	(0.21, 0.90, 0.34)
2 LSCs	(0.32, 0.13, 0.19)	(0.16, 0.87, 0.27)

Note. The GC test set contains 288 sources.

minimum number of samples required to split an internal node, the minimum number of samples required to be at a leaf node, and the number of features to consider when looking for the best split. For a detailed description of the inner workings of random forest algorithms, see G. Biau & E. Scornet (2015). We also tested, as a preprocessing step, undersampling the majority classes (the contaminants), oversampling the minority classes (the confirmed GCs), and combinations of both—that is, respectively excluding and/or imputing data points to mitigate potential biases due to class imbalance when training the models. These procedures did not change the performance of the models, compared to cases where the number of members of each class was not altered but stratified CV was used. Hence we decided not to use under-/oversampling methods.

In the end, the best-performing models in terms of the F1-score for the GC class were evaluated on the test subset (20% of the catalog) of each run.

4. Results

We ran PCA and AEs over our LSST-like photometric catalog to transform the colors into PCs and LSCs and assess how distinguishable extragalactic, point-like GCs are from the background galaxies and foreground stars using two classification models and eight input sets. Table 5 displays, for the GC class, the performance metrics (precision, recall, F1-score) of the best models for each input type.

The first direct result that can be extracted from Table 5 is that the best precision score for the GC class, 71%, was obtained by RFC, using 15 PCs as input, with a respective 29% recall rate (as 71% of the actual GCs in the test set are not identified by the model). Given that the test set was selected while preserving the original class proportions, we can view the evaluation of each model on the test set as a GC candidate selection process. In that sense, the performance of the RFC using 15 PCs as input corresponds to a sample of candidates, 29% contaminated and only 29% complete. At the same time, the performance of the RFCs that received 15 colors, 15 PCs, and 4 PCs as input is almost identical, with a 2% difference in precision to make the 15 PCs case stand out, which is hardly statistically significant. From the perspective of the MLPCs, with more homogeneous performance scores across the various inputs, recall rates of ∼90% are obtained, although associated with even lower precision and, in some cases, similar F1-scores compared to the other models with the same inputs. That is, with MLPCs, despite contamination rates of ∼80%, the samples of GC candidates would contain ∼90% of the actual GCs. The performance metrics for the classification of the background galaxies and foreground stars (the two other classes) are presented in Appendix B as supplementary results.

Figure 7 presents the confusion matrices (CMs) for the RFC and MLPC that receive 15 PCs as input (the fourth row of Table 5); the absolute numbers in the matrices allow a more direct and detailed assessment of the precision and recall scores discussed previously. This RFC misclassifies 71% of the 288 actual GCs in the test sample, of which 69% (198) are labeled galaxies and 2% (5) are labeled stars (the bottom row of the blue CM). Although it correctly identifies 29% of the actual GCs (85), it also misclassifies 33 galaxies and 2 stars as GCs, resulting in 35 contaminants and an overall precision of 71% (the right column of the blue CM). Again, this precision score is practically the same as the one achieved by the RFC with the 15-color input. From the perspective of the MLPC, in comparison with the corresponding RFC, galaxies are less confused with actual GCs (the bottom row of the red CM), but actual galaxies and stars are more confused with GCs (the right column of the red CM). Quantitatively, if the right columns of the two CMs are proportionally compared, the RFC is demonstrated to be more accurate when selecting GCs: MLPC misclassifies about 3 times more galaxies and stars as GCs than the RFC does, despite correctly identifying 89% of the actual GCs—hence its lower F1-score of 35%. The CMs associated with all other runs are available in Appendix B.

In terms of the F1-scores for the GC class, the best-performing RFCs are systematically superior to their MLPC counterparts. This, together with the provided description of the CMs, suggests a limitation inherent to the dataset. More complex models may not improve the performance on the same dataset. Therefore, a concrete limit to the contribution of the color space to the selection of point-like extragalactic GC candidates is obtained from our LSST-like catalog: a very incomplete sample of candidates (recall <30%) may yield a minimum contamination rate of ∼30%, such that increasing completeness may lead to higher contamination.

It is also possible to demonstrate that PCA has some compression efficacy for this dataset; this is supported by the fact that the RFCs with 15 colors and 15 PC inputs show the same performance scores as those from the RFC with 4 PC inputs, while the RFC with 4 color inputs reported a substantially lower precision score for the GC class. In contrast, the RFC with 2 color inputs yields a similar, or slightly superior, performance compared to the 2 PC input RFC. Furthermore, the RFC that uses 4 LSCs as input also outperforms the one that uses 4 colors, but it is equivalent (or slightly inferior) to the one that uses 4 PCs as input. Notably, all 2D inputs yield substantially lower precision scores compared to their higher-dimensional counterparts, including the cases of PCs and LSCs, which encode information from more than two colors. This specific outcome highlights the limitation of 2D color–color diagrams, and other 2D projections, to select GC candidates compared to the use of higher-dimensional projections. Finally, the fact that the models that received LSCs do not outperform those associated with PCs indicates that there is no nonlinear relation within the color space of this dataset that could be used to reduce contamination. This is quantitative evidence in accordance with the suspicion raised by visual inspection of the 2D projects in Figure 6.

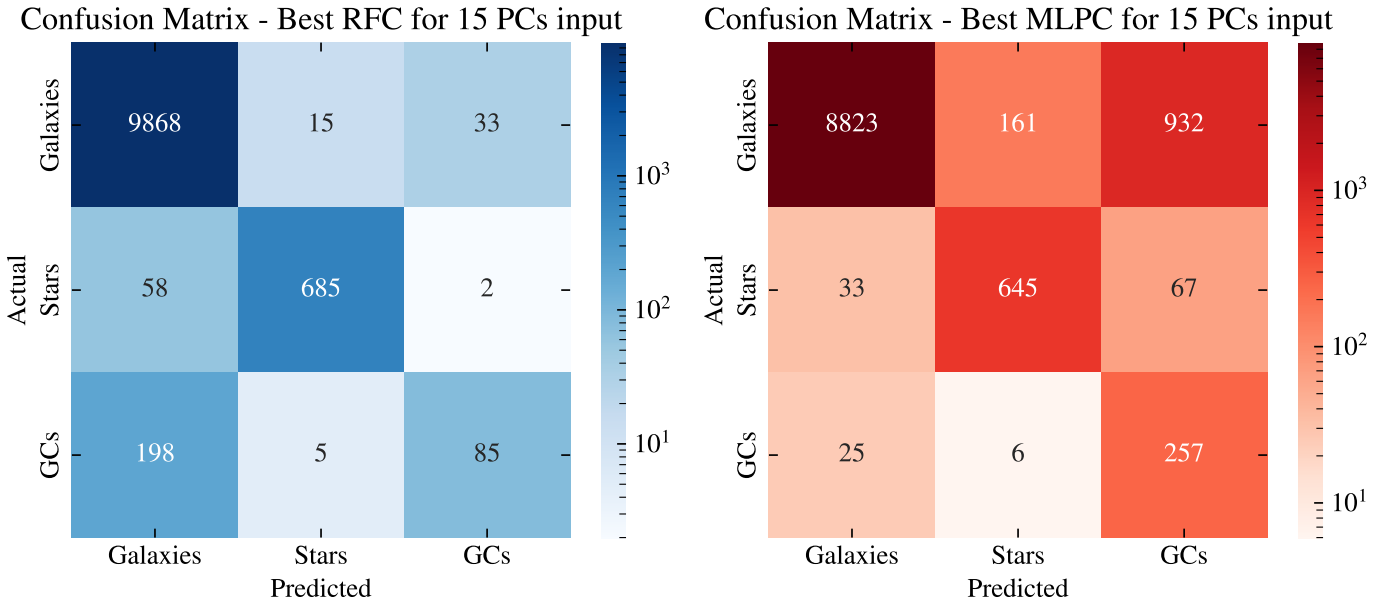

Figure 7. The CMs of the RFC and MLPC that received the set of all 15 PCs as input. These are also the best-performing models in terms of both precision and F1-score for the GC class, although they behave almost identically to the ones that received 15 colors as input.

5. Discussions and Conclusions

We have assembled an LSST-like (FDS + DES) photometric catalog of the central region of the Fornax Cluster and prepared labels for confirmed GCs from the literature, and background galaxies and foreground stars selected from DES DR2 data. Our goal was to understand to what extent the color space of this catalog enables us to correctly identify extragalactic, point-like GCs among contaminants. Using RFCs, we show that projecting the catalog colors onto their PCs allows for dimensionality reduction without compromising the precision of GC identification. Namely, the minimum contamination rate of ∼30% is unchanged, regardless if one uses all 15 available colors or only the first 4 PCs of these colors as input to RFCs. Nevertheless, such a contamination level is achieved at the expense of highly incomplete GC candidate selection. MLPCs do not yield improved performance, indicating an intrinsic limitation of the data. It was also possible to show the use of LSCs from nonlinear AEs yielded equivalent or less accurate results when compared to the use of PCs and colors.

This leads us to encourage the use of PCs of colors instead of colors themselves when selecting extragalactic GC candidates, especially in scenarios where many photometric bands are available. For instance, the ground-based Javalambre Physics of the Accelerating Universe Astrophysical Survey (J-PAS) uses a set of broad, intermediate, and narrowband filters (59 in total), producing SEDs with more encoded information about the nature of the objects compared to *ugrizY* SEDs; the color space of a J-PAS photometric catalog is composed of ∼1000 dimensions. A pipeline to extract samples of extragalactic GCs from such a high-dimensional dataset could use PCA and RFCs as a solid starting point, although it is expected that more complex models could indeed be useful in this case.

The limited ability of *ugrizY* colors to discriminate between stars, galaxies, and GCs is perhaps not too surprising. In general, the ultraviolet-to-near-infrared emission of galaxies and star clusters is dominated by starlight. Galaxy light can also include emission from active galactic nuclei and both absorption and emission from the interstellar medium, and is (of course) affected by redshifting. The colors of the simple stellar populations of GCs change with age as different stellar evolutionary stages dominate, but galaxies also contain stars with a range of ages. That is, completely distinguishing between the simpler star formation histories of GCs and the more complex ones of galaxies is not possible with *ugrizY* photometry alone. In fact, similar data-driven investigations distinguishing extragalactic GCs from contaminants have found that morphometric quantities may hold the most discriminatory information, except when attempting to separate the faintest/smallest GCs from foreground stars. For instance, M. Mohammadi et al. (2022) also employ a supervised learning setup to a set of confirmed GCs and ultracompact dwarf galaxies, background galaxies, and foreground stars in the Fornax Cluster, using the colors and FWHMs from the *ugriJK*$_s$ (FDS + VISTA) filter set. They found that the optical FWHMs were more important for separating background galaxies from GCs than any other color, apart from $g - i$ and $g - r$ (which have the highest S/N), while using a more restricted sample of bright GCs compared to this work. A similar effect is observed in E. Barbisan et al. (2022), which uses *ugriz* colors, magnitudes, and flux radii data of M87 from the Next Generation Virgo Cluster Survey to construct a model to select GC candidates, again, in a supervised setup. They found optical flux radii to be more informative than colors, although they use a very restricted sample of 90 extended background galaxies, together with 1160 bright spectroscopically confirmed GCs and 2346 foreground stars.

Therefore, additional steps to continue reducing contamination in samples of extragalactic GC candidates for multiband surveys like LSST must rely on complementary, more informative data to augment the color space before attempting to leverage more complex models. Possibilities include, most importantly, morphometric properties, which are known to be very effective discriminators when sufficient spatial resolution is achieved, as in spaced-based facilities such as HST, Euclid,

and the upcoming Roman Space Telescope (E. W. Peng et al. 2006; Euclid Collaboration et al. 2025; J. M. Howell et al. 2025; S. S. Larsen et al. 2025); near-infrared, as demonstrated by R. P. Muñoz et al. (2014) and M. Cantiello et al. (2018) to be very useful; astrometric parallax from Gaia (K. T. Voggel et al. 2020; A. L. Chies-Santos et al. 2022); and careful ultraviolet contribution (e.g., T. A. Pacheco et al. 2025). Specifically regarding the scenario in which space-based imaging in multiple filters is available, the use of convolutional neural networks and methods alike to identify very promising extragalactic GC candidates from the images themselves, or from the image cutouts of photometric candidates, becomes not only feasible but perhaps desirable, especially for automated selection over large areas of the sky (D. Dold & K. Fahrion 2022).

With that in mind, we highlight the importance of collective work to fuse the scientific potentials of different facilities/surveys/collaborations and thus foster the perspectives of extragalactic GC science. For example, the first major effort to perform joint analysis involving Roman and Rubin/LSST data was "OpenUniverse2024" (OpenUniverse et al. 2025), which produced 4 million simulated individual images covering two overlapping areas: an approximate 70° field, to be observed by both the LSST Wide-Fast-Deep Survey and the Roman High-Latitude Wide-Area Survey, and the LSST ELAIS-S1 Deep-Drilling Field, which is also to be observed by the Roman High-Latitude Time-Domain Survey. A large collaboration of NASA, the US National Science Foundation (NSF), and the US Department of Energy (DOE) was involved in the "OpenUniverse2024" Project (which specifically included the NASA OpenUniverse team, the LSST Dark Energy Science Collaboration, the Roman High-Latitude Imaging Survey Project Infrastructure Team, and the Roman Supernova Project Infrastructure Team, as well as several other scientists). Another major joint analysis effort is underway by a Roman Wide Field Science Team. They will develop the Scarlet2 software package (a multiband, multiresolution astronomical source modeling framework to perform joint detection, deblending/modeling, and measurement for Roman and Rubin data). Catalog data products from this work are expected to be released in 2027 and 2028.

Acknowledgments

The authors thank the reviewer of this paper, whose comments and suggestions greatly enriched the quality of this work. N.S.S. acknowledges support from Laboratório Interinstitucional de e-Astronomia (LIneA, Brazil), along with the Brazilian agencies Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação Araucária, and Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS). A.C.S. acknowledges support from FAPERGS (grant Nos. 23/2551-0001832-2 and 24/2551-0001548-5), CNPq (grant Nos. 314301/2021-6, 312940/2025-4, 445231/2024-6, and 404233/2024-4), and CAPES (grant No. 88887.004427/2024-00). R.S.S. acknowledges support from CNPq (grant Nos. 446508/2024-1 and 315026/2025-1). M.C. acknowledges support from ASI–INAF grant No. 2024-10-HH.0 (WP8420), the ESO Scientific Visitor Program, and INAF GO-grant No. 12/2024 (P.I. M. Cantiello). T.S. acknowledges funding from the CNES postdoctoral fellowship program. J.P.C. acknowledges support from Consejo Nacional de Investigaciones Científicas y Técnicas de la República Argentina, Agencia Nacional de Promoción Científica y Tecnológica, and Universidad Nacional de La Plata (Argentina).

Author Contributions

Conceptualization, A.C.S., R.S.S., K.D., N.S.S., J.P.C., C.B.; methodology, R.S.S., A.C.S., N.S.S.; data curation, N.S.S.; software, N.S.S., R.S.S.; formal analysis, N.S.S.; visualization, N.S.S.; resources, A.C.S., C.B.; supervision, A.C.S., R.S.S., J.P.C., C.B., T.P., K.D., M.C., R.M.; writing (original draft preparation), N.S.S.; writing (review and editing), N.S.S., A.C.S., K.D., K.R., P.B., M.C., J.S., A.I.E., J.P.C., T.S., T.P., P.S.L., P.F., R.M., Y.O.B., J.G., N.P.; project administration, A.C.S., K.D., N.S.S.; funding acquisition, A.C.S., J.P.C., R.M.

Appendix A
Comparison of FDS and DES Photometric Data

Figures 8 and 9 allow us to draw the same conclusion as obtained from the inspection of Figure 3: `MAG_APER_5` is the circular aperture model used to perform photometry on DES images that most closely resembles FDS PSF photometry.

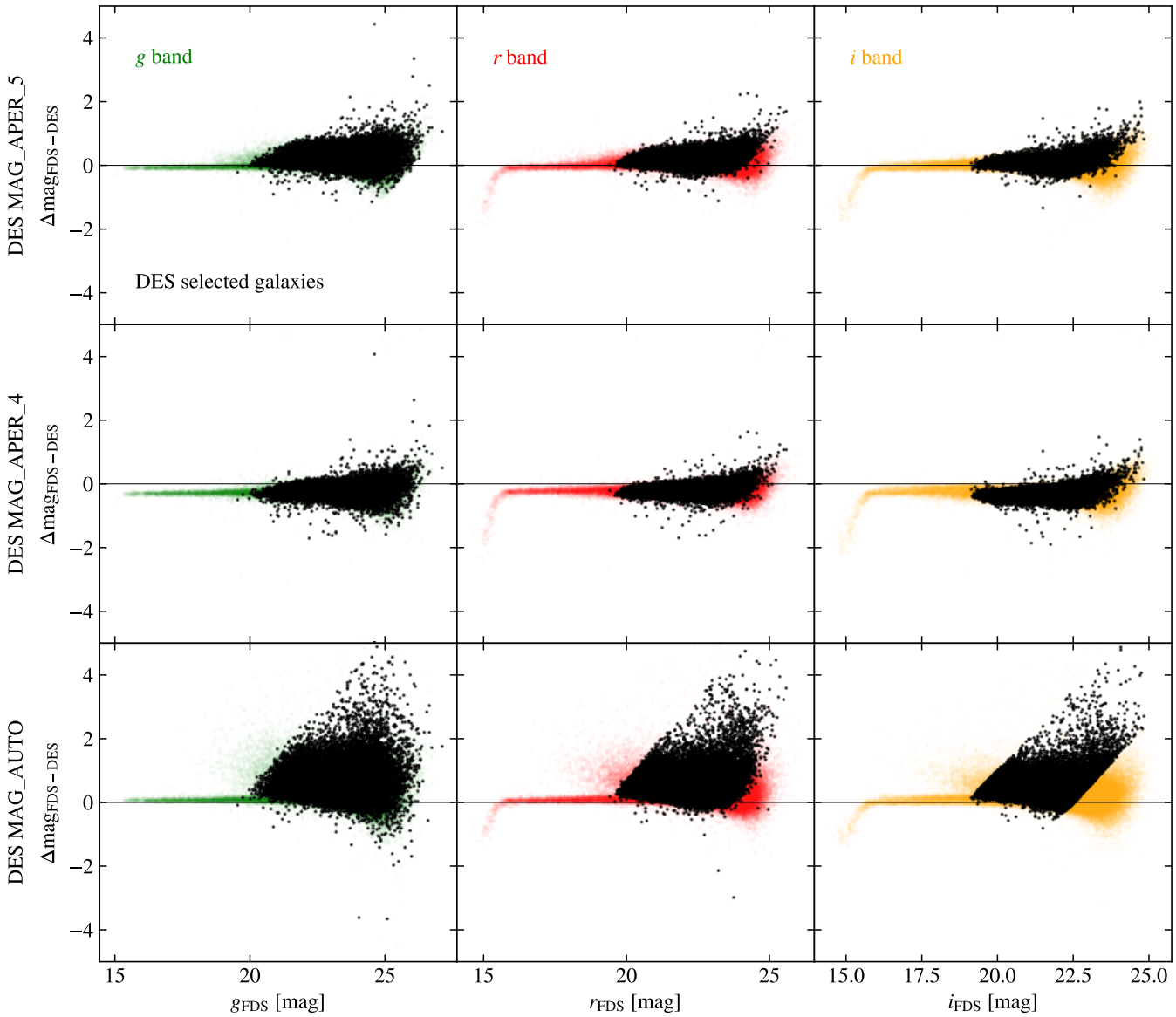

Figure 8. $\Delta\mathrm{mag}_{\mathrm{FDS-DES}}$ vs. g, r, i_{FDS}: the difference in magnitude for the same source, in the same band, but in different surveys against the FDS magnitude in the same band. The first row displays the plots where DES `MAG_APER_5` ($2\farcs92$) data were used, the second DES `MAG_APER_4` ($1\farcs92$), and the third DES `MAG_AUTO`. Black points denote background galaxies selected via Criterion (1), as in Section 2.2. The horizontal black line is $y = 0$.

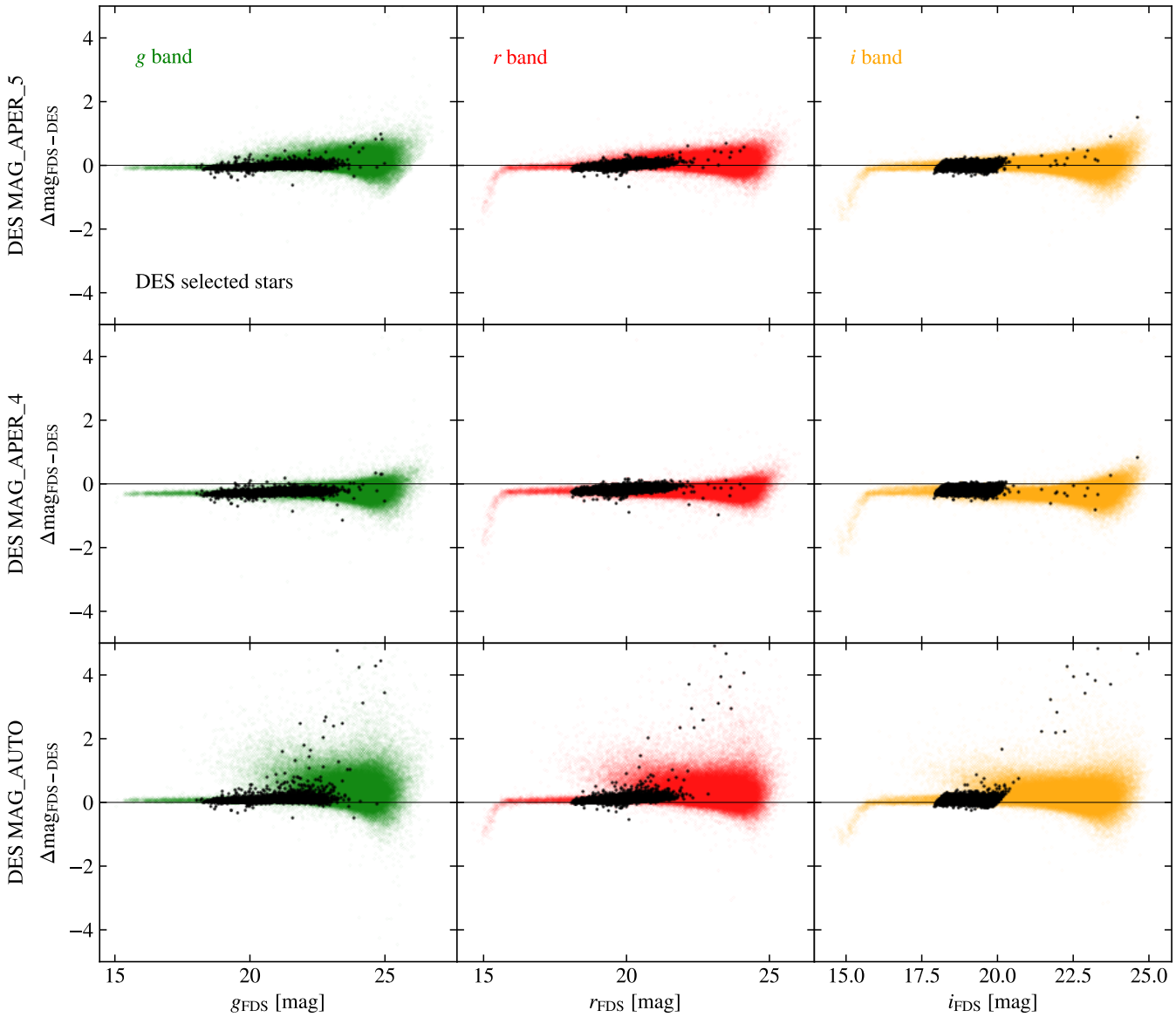

Figure 9. $\Delta\mathrm{mag}_{\mathrm{FDS-DES}}$ vs. g, r, i_{FDS}: the difference in magnitude for the same source, in the same band, but in different surveys against the FDS magnitude in the same band. The first row displays the plots where DES `MAG_APER_5` (2.″92) data were used, the second DES `MAG_APER_4` (1.″92), and the third DES `MAG_AUTO`. Black points denote foreground stars selected via Criterion (2), as in Section 2.2. The horizontal black line is $y = 0$.

Appendix B
Supplementary Results

In Tables 6 and 7, we present the models' performance scores on the classification of background galaxies and foreground stars, respectively. Figures 10–16 are the CMs of all other model runs, apart from those shown in Figure 7.

Table 6
Results for the Classification of Background Galaxies: Output Metric Triplets (Precision, Recall, F1-score) of the Best-performing Models (the Ones Whose Hyperparameter Values Maximized the F1-score for the Galaxy Class)

	RFC	MLPC
15 colors	(0.97, 0.99, 0.98)	(0.99, 0.87, 0.93)
4 colors	(0.97, 0.99, 0.98)	(0.99, 0.86, 0.92)
2 colors	(0.96, 0.98, 0.97)	(0.99, 0.83, 0.90)
15 PCs	(0.97, 0.99, 0.98)	(0.99, 0.88, 0.93)
4 PCs	(0.97, 0.99, 0.98)	(0.99, 0.87, 0.93)
2 PCs	(0.95, 0.98, 0.97)	(0.99, 0.82, 0.90)
4 LSCs	(0.97, 0.99, 0.98)	(0.99, 0.87, 0.93)
2 LSCs	(0.95, 0.98, 0.97)	(0.99, 0.84, 0.91)

Note. The galaxy test set contains 9916 sources.

Table 7
Results for the Classification of Foreground Stars: Output Metric Triplets (Precision, Recall, F1-score) of the Best-performing Models (the Ones Whose Hyperparameter Values Maximized the F1-score for the Star Class)

	RFC	MLPC
15 colors	(0.96, 0.91, 0.93)	(0.64, 0.89, 0.74)
4 colors	(0.91, 0.85, 0.88)	(0.60, 0.84, 0.70)
2 colors	(0.75, 0.70, 0.72)	(0.49, 0.79, 0.60)
15 PCs	(0.97, 0.92, 0.94)	(0.67, 0.86, 0.75)
4 PCs	(0.94, 0.91, 0.92)	(0.71, 0.81, 0.76)
2 PCs	(0.76, 0.67, 0.71)	(0.49, 0.82, 0.62)
4 LSCs	(0.93, 0.89, 0.91)	(0.61, 0.86, 0.71)
2 LSCs	(0.75, 0.67, 0.71)	(0.53, 0.70, 0.60)

Note. The star test set contains 745 sources.

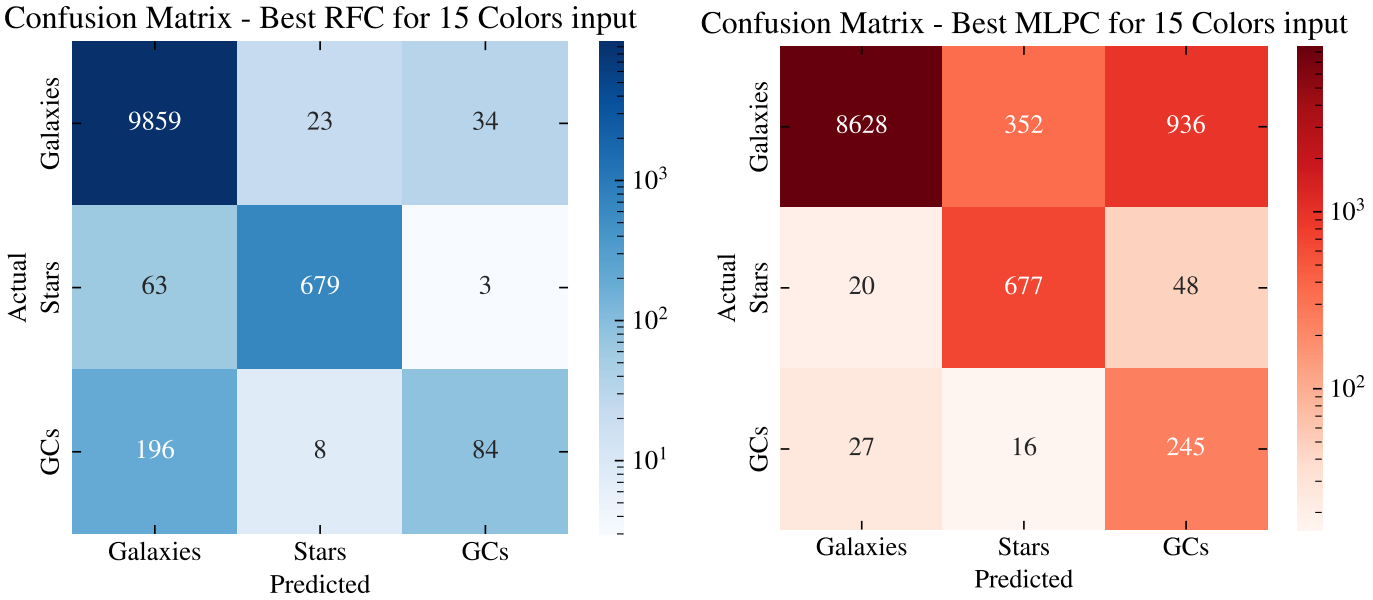

Figure 10. The CMs of the RFC and MLPC that received 15 colors as input.

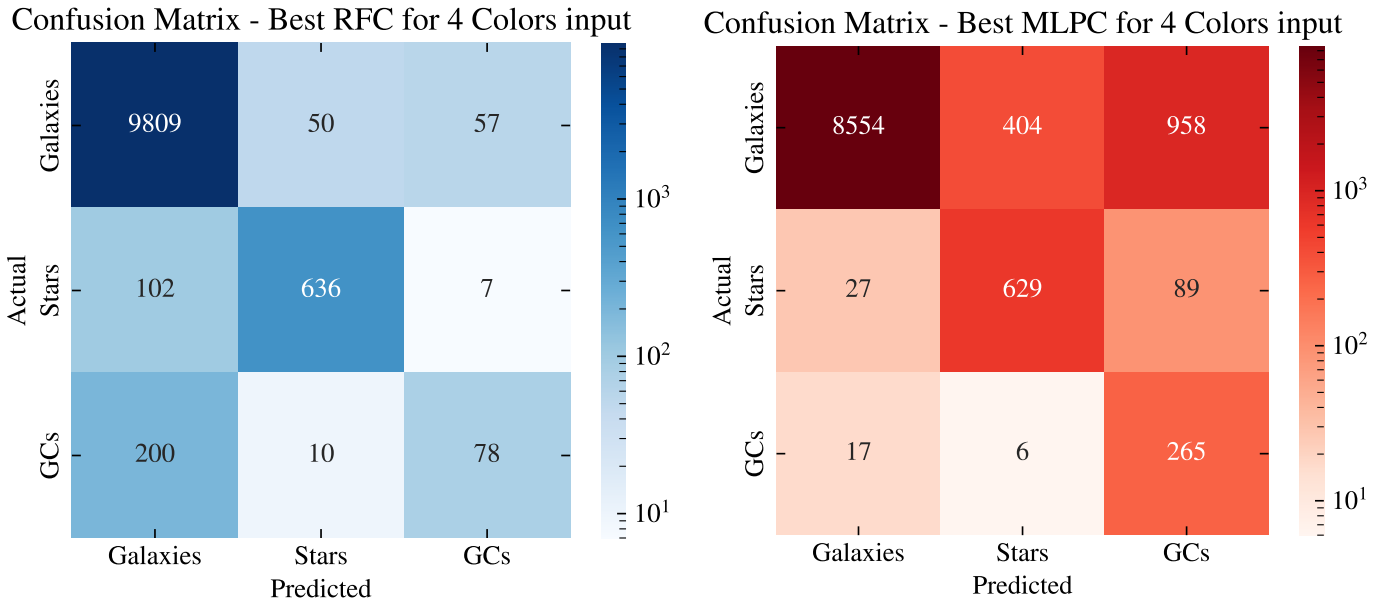

Figure 11. The CMs of the RFC and MLPC that received 4 colors as input.

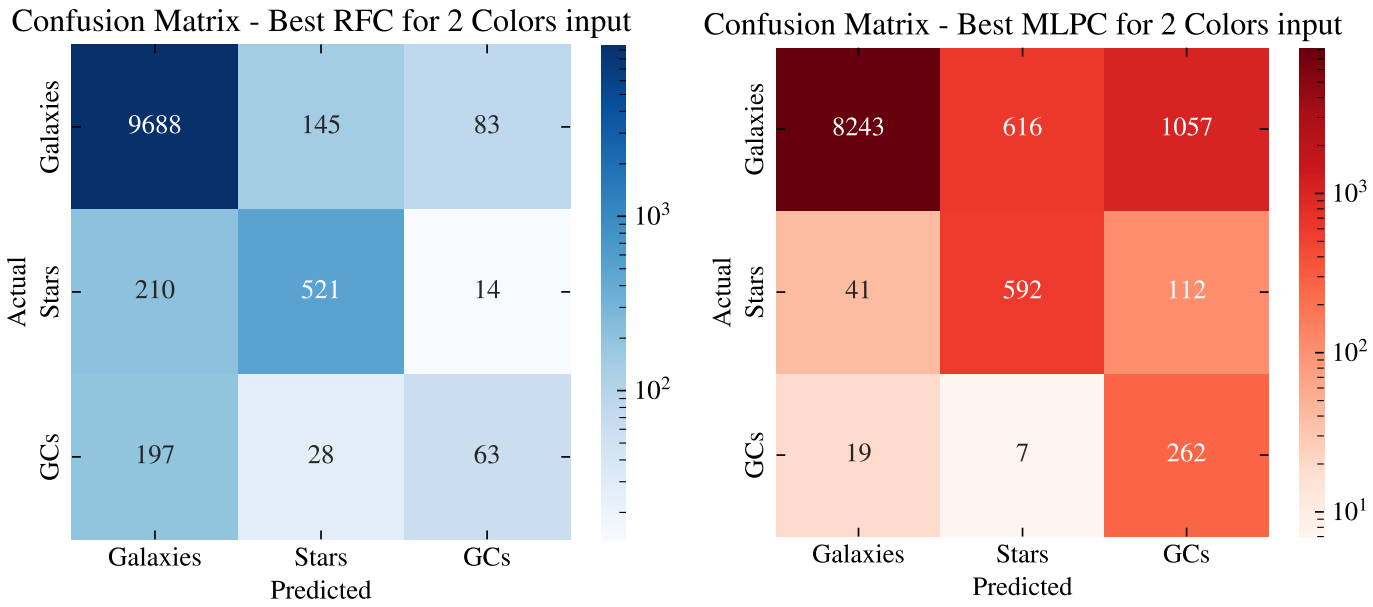

Figure 12. The CMs of the RFC and MLPC that received 2 colors as input.

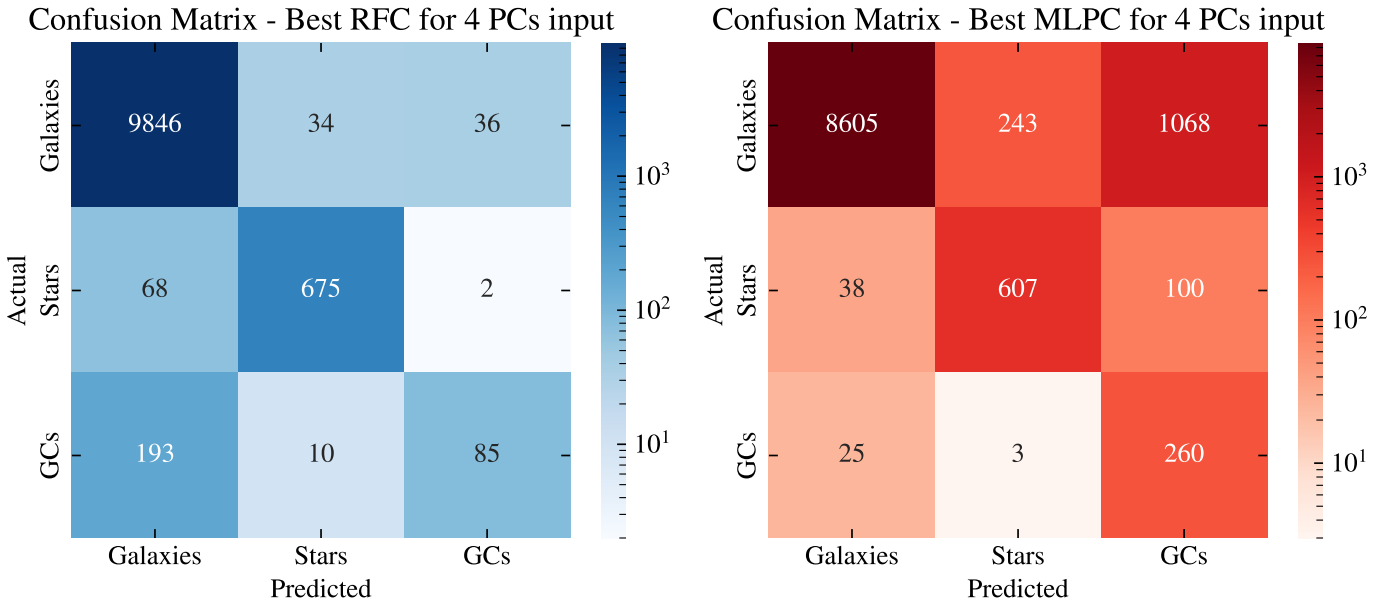

Figure 13. The CMs of the RFC and MLPC that received 4 PCs as input.

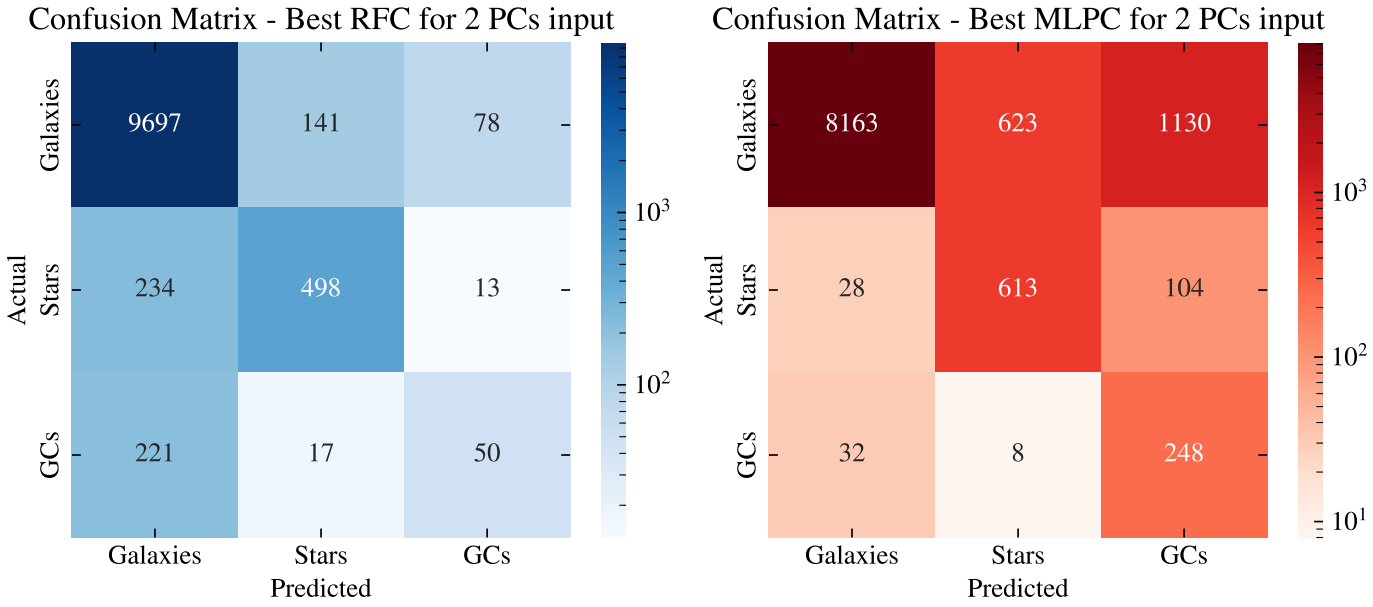

Figure 14. The CMs of the RFC and MLPC that received 2 PCs as input.

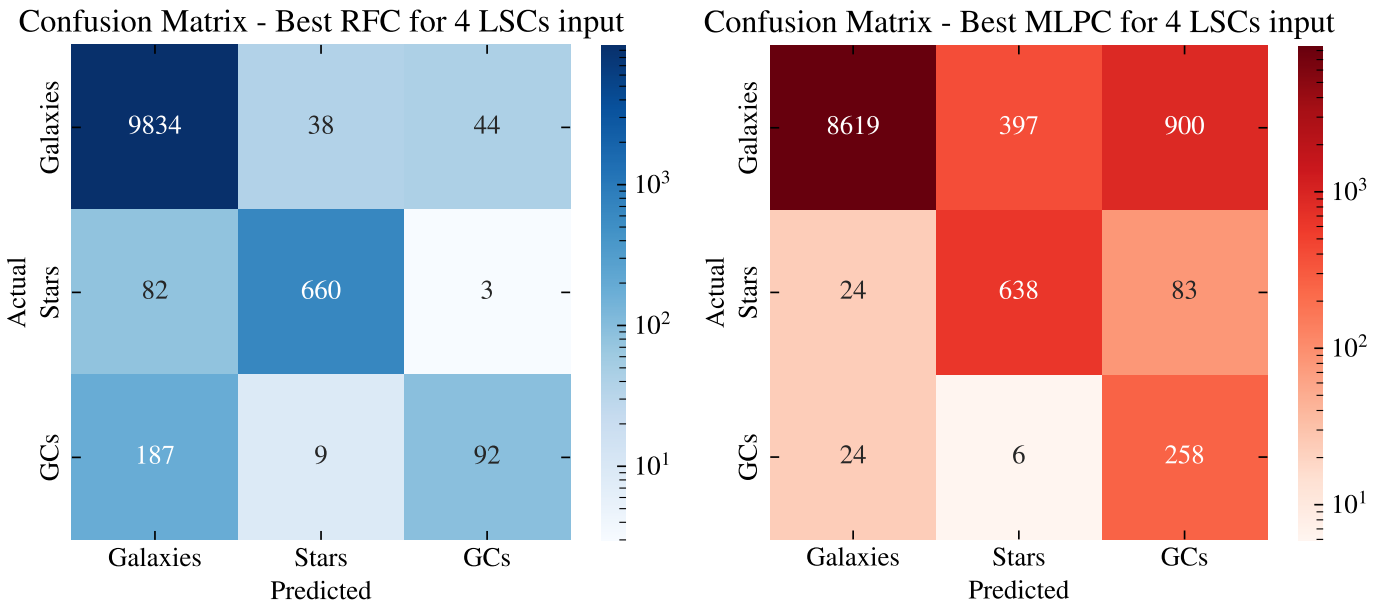

Figure 15. The CMs of the RFC and MLPC that received 4 LSCs as input.

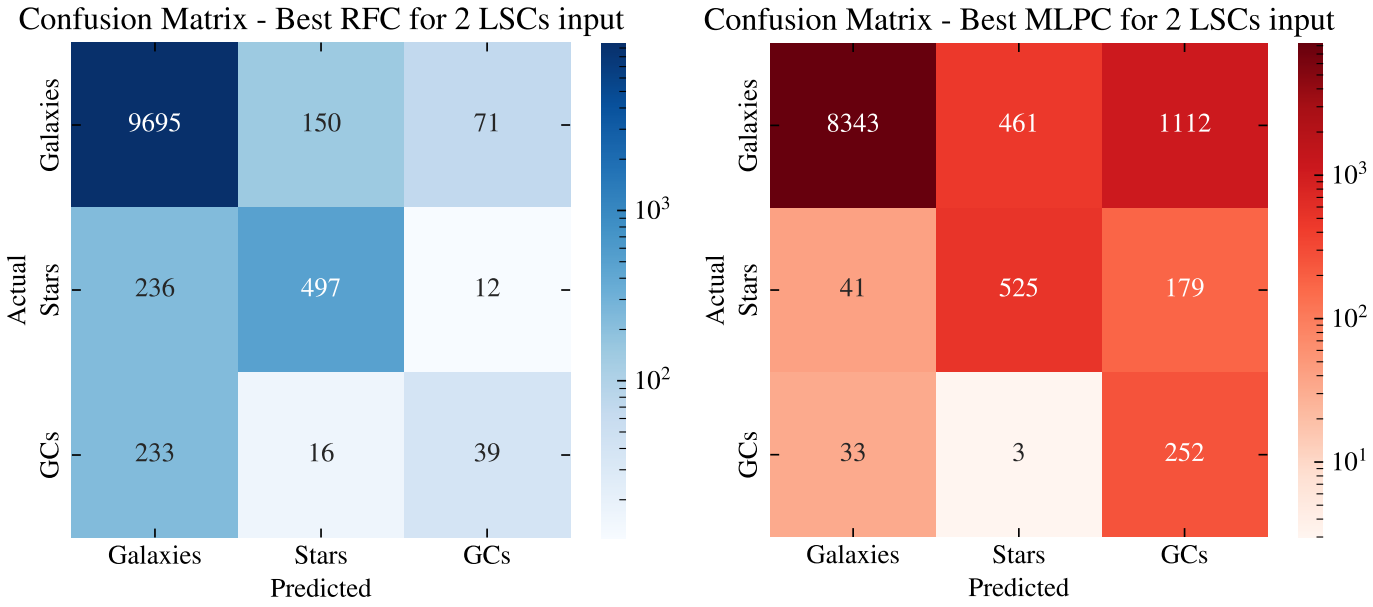

Figure 16. The CMs of the RFC and MLPC that received 2 LSCs as input.

ORCID iDs

Nicholas Schweder-Souza https://orcid.org/0009-0001-0407-8134
Ana L. Chies-Santos https://orcid.org/0000-0003-3220-0165
Rafael S. de Souza https://orcid.org/0000-0001-7207-4584
Kristen C. Dage https://orcid.org/0000-0002-8532-4025
Charles J. Bonatto https://orcid.org/0000-0002-4102-1751
Juan P. Caso https://orcid.org/0000-0003-0812-9928
Michele Cantiello https://orcid.org/0000-0003-2072-384X
Pedro dos Santos-Lopes https://orcid.org/0009-0005-7299-4168
Pedro Floriano https://orcid.org/0009-0008-4034-7670
Thayse A. Pacheco https://orcid.org/0000-0002-8139-7278
Katherine L. Rhode https://orcid.org/0000-0001-8283-4591
Pauline Barmby https://orcid.org/0000-0003-2767-0090
Jennifer Sobeck https://orcid.org/0000-0002-4989-0353
Ana I. Ennis https://orcid.org/0000-0001-8411-8783
Yasna Ordenes-Briceño https://orcid.org/0000-0001-7966-7606
Teymoor Saifollahi https://orcid.org/0000-0002-9554-7660
Julia Gschwend https://orcid.org/0000-0003-3023-8362
Niranjana P. https://orcid.org/0000-0003-3372-3638
Rubens E. G. Machado https://orcid.org/0000-0001-7319-297X

References

Abbott, T. M. C., Abdalla, F. B., Allam, S., et al. 2018, ApJS, 239, 18
Abbott, T. M. C., Adamów, M., Aguena, M., et al. 2021, ApJS, 255, 20
Adamo, A., Usher, C., Pfeffer, J., & Claeyssens, A. 2023, MNRAS, 525, L6
Aihara, H., Armstrong, R., Bickerton, S., et al. 2018, PASJ, 70, S8
Akeson, R., Armus, L., Bachelet, E., et al. 2019, arXiv:1902.05569
Anand, G. S., Tully, R. B., Cohen, Y., et al. 2024, ApJ, 973, 83
Annibali, F., Morandi, E., Watkins, L. L., et al. 2018, MNRAS, 476, 1942
Austin, P. C. 2011, Multivariate Behav Res, 46, 399
Bank, D., Koenigstein, N., & Giryes, R. 2023, Autoencoders, Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook (Springer), 353
Barbisan, E., Huang, J., Dage, K. C., et al. 2022, MNRAS, 514, 943
Baumgardt, H., & Hilker, M. 2018, MNRAS, 478, 1520
Beasley, M. A. 2020, in Reviews in Frontiers of Modern Astrophysics; From Space Debris to Cosmology, ed. P. Kabáth, D. Jones, & M. Skarka (Springer International), 245
Berkheimer, J. M., Windhorst, R. A., Harris, W. E., et al. 2025, AJ, 171, 48
Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393
Bianco, F. B., Ivezić, Ž., Jones, R. L., et al. 2022, ApJS, 258, 1
Biau, G., & Scornet, E. 2015, TEST, 25, 197
Breiman, L. 2001, MachL, 45, 5
Brodie, J. P., Romanowsky, A. J., Strader, J., & Forbes, D. A. 2011, AJ, 142, 199
Brodie, J. P., Romanowsky, A. J., Strader, J., et al. 2014, ApJ, 796, 52
Burkert, A., & Forbes, D. A. 2020, AJ, 159, 56
Canossa-Gosteinski, M. A., Chies-Santos, A. L., Furlanetto, C., et al. 2024, MNRAS, 534, 1729
Cantiello, M., Grado, A., Rejkuba, M., et al. 2018, A&A, 611, A21
Cantiello, M., Venhola, A., Grado, A., et al. 2020, A&A, 639, A136
Caso, J. P., Bassino, L. P., Richtler, T., Calderón, J. P., & Smith Castelli, A. V. 2014, MNRAS, 442, 891
Chaturvedi, A., Hilker, M., Cantiello, M., et al. 2022, A&A, 657, A93
Chies-Santos, A. L., de Souza, R. S., Caso, J. P., et al. 2022, MNRAS, 516, 1320
Chies-Santos, A. L., Larsen, S. S., Kuntschner, H., et al. 2011a, A&A, 525, A20
Chies-Santos, A. L., Larsen, S. S., Wehner, E. M., et al. 2011b, A&A, 525, A19
Côté, P., Blakeslee, J. P., Ferrarese, L., et al. 2004, ApJS, 153, 223
CSST Collaboration, Gong, Y., Miao, H., et al. 2025, SCPMA, 69, 239501
D'Abrusco, R., Cantiello, M., Paolillo, M., et al. 2016, ApJL, 819, L31
Dage, K. C., Zepf, S. E., Thygesen, E., et al. 2020, MNRAS, 497, 596
Dal Tio, P., Pastorelli, G., Mazzi, A., et al. 2022, ApJS, 262, 22
de Souza, R. S., Maio, U., Biffi, V., & Ciardi, B. 2014, MNRAS, 440, 240
de Souza, R. S., Quanfeng, X., Shen, S., Peng, C., & Mu, Z. 2022, A&C, 41, 100633
Diego, J. M., Pascale, M., Frye, B., et al. 2023, A&A, 679, A159
Dold, D., & Fahrion, K. 2022, A&A, 663, A81
Dornan, V., & Harris, W. E. 2025, ApJ, 988, 70
Dou, H., Li, H., Zhang, H., Yu, H., & Wang, H. 2025, ApJ, 994, 96
Euclid Collaboration, Scaramella, R., Amiaux, J., et al. 2022, A&A, 662, A112
Euclid Collaboration, Voggel, K., Lançon, A., et al. 2025, A&A, 693, A251
Fahrion, K., Lyubenova, M., Hilker, M., et al. 2020, A&A, 637, A27
Ferrarese, L., Côté, P., Cuillandre, J.-C., et al. 2012, ApJS, 200, 4
Forbes, D. A., Ferré-Mateu, A., Gannon, J. S., et al. 2022, MNRAS, 512, 802
Forbes, D. A., Read, J. I., Gieles, M., & Collins, M. L. M. 2018, MNRAS, 481, 5592
Fournier, Q., & Aloise, D. 2019, in IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (IEEE), 211
Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., et al. 2016, A&A, 595, A1

Garcia-Dias, R., Vieira, S., Lopez Pinaya, W. H., & Mechelli, A. 2020, in Machine Learning, ed. A. Mechelli & S. Vieira (Academic Press), 227
Girardi, L., Groenewegen, M. A. T., Hatziminaoglou, E., & da Costa, L. 2005, A&A, 436, 895
Grasser, N., Arnaboldi, M., Barbosa, C. E., et al. 2024, A&A, 683, A8
Hargis, J. R., & Rhode, K. L. 2012, AJ, 144, 164
Harris, W. E., Harris, G. L. H., & Alessi, M. 2013, ApJ, 772, 82
Harris, W. E., Morningstar, W., Gnedin, O. Y., et al. 2014, ApJ, 797, 128
Ho, D. E., Imai, K., King, G., & Stuart, E. A. 2007, Political Analysis, 15, 199
Howell, J. M., Ferguson, A. M. N., Larsen, S. S., et al. 2025, A&A, 706, A185
Hudson, M. J., Harris, G. L., & Harris, W. E. 2014, ApJL, 787, L5
Hughes, A. K., Sand, D. J., Seth, A., et al. 2021, ApJ, 914, 16
Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111
Jolliffe, I. T., & Cadima, J. 2016, RSPTA, 374, 20150202
Jordán, A., Blakeslee, J. P., Côté, P., et al. 2007, ApJS, 169, 213
Jordán, A., Peng, E. W., Blakeslee, J. P., et al. 2009, ApJS, 180, 54
Jordán, A., Peng, E. W., Blakeslee, J. P., et al. 2015, ApJS, 221, 13
Kirsten, F., Marcote, B., Nimmo, K., et al. 2022, Natur, 602, 585
Kuhn, M. A., de Souza, R. S., Krone-Martins, A., et al. 2021, ApJS, 254, 33
Larsen, S. S., Ferguson, A. M. N., Howell, J. M., et al. 2025, A&A, 703, A113
Lim, S., Peng, E. W., Côté, P., et al. 2025, ApJS, 276, 34
Lomelí-Núñez, L., Mayya, Y. D., Rodríguez-Merino, L. H., et al. 2024, MNRAS, 528, 1445
Maccarone, T. J., Kundu, A., Zepf, S. E., & Rhode, K. L. 2007, Natur, 445, 183
Masters, K. L., Jordán, A., Côté, P., et al. 2010, ApJ, 715, 1419
Mirabile, M., Cantiello, M., Lonare, P., et al. 2024, A&A, 691, A104
Mohammadi, M., Mutatiina, J., Saifollahi, T., & Bunte, K. 2022, A&C, 39, 100555
Muñoz, R. P., Puzia, T. H., Lançon, A., et al. 2014, ApJS, 210, 4
Murtagh, F. 1991, Neurocomputing, 2, 183
Nikutta, R., Fitzpatrick, M., Scott, A., & Weaver, B. A. 2020, A&C, 33, 100411
Ochsenbein, F., Bauer, P., & Marcout, J. 2000, A&AS, 143, 23
OpenUniverseLSST Dark Energy Science CollaborationRoman HLIS Project Infrastructure, et al. 2025, MNRAS, 544, 3799
Pacheco, T. A., Coelho, P. R. T., Martins, L. P., et al. 2025, ApJ, 992, 151
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, JMLR, 12, 2825
Peletier, R., Iodice, E., Venhola, A., et al. 2020, arXiv:2008.12633
Peng, E. W., Jordán, A., Côté, P., et al. 2006, ApJ, 639, 95
Pota, V., Napolitano, N. R., Hilker, M., et al. 2018, MNRAS, 481, 1744
Reina-Campos, M., Trujillo-Gomez, S., Pfeffer, J. L., et al. 2023, MNRAS, 521, 6368
Rejkuba, M. 2012, Ap&SS, 341, 195
Saifollahi, T., Janz, J., Peletier, R. F., et al. 2021, MNRAS, 504, 3580
Saifollahi, T., Lançon, A., Cantiello, M., et al. 2025b, A&A, 703, A184
Saifollahi, T., Voggel, K., Lançon, A., et al. 2025a, A&A, 697, A10
Saifollahi, T., Zaritsky, D., Trujillo, I., et al. 2022, MNRAS, 511, 4633
Schlafly, E. F., & Finkbeiner, D. P. 2011, ApJ, 737, 103
Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, ApJ, 500, 525
Schuberth, Y., Richtler, T., Hilker, M., et al. 2010, A&A, 513, A52
Usher, C., Brodie, J. P., Forbes, D. A., et al. 2019, MNRAS, 490, 491
Usher, C., Caldwell, N., & Cabrera-Ziri, I. 2024, MNRAS, 528, 6010
Usher, C., Dage, K. C., Girardi, L., et al. 2023, PASP, 135, 074201
Valenzuela, L. M., Moster, B. P., Remus, R.-S., O'Leary, J. A., & Burkert, A. 2021, MNRAS, 505, 5815
Voggel, K. T., Seth, A. C., Sand, D. J., et al. 2020, ApJ, 899, 140
Xu, Q., Shen, S., de Souza, R. S., et al. 2023, MNRAS, 526, 6391
Zaritsky, D. 2022, MNRAS, 513, 2609