Detection of extragalactic Ultra-Compact Dwarfs and Globular Clusters using Explainable AI techniques

Mohammad Mohammadi[a,1,*], Jarvin Mutatiina[a,1], Teymoor Saifollahi[b,1], Kerstin Bunte[a]

Faculty of Science and Engineering, University of Groningen, The Netherlands

[a]*Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence*
[b]*Kapteyn Astronomical Institute, University of Groningen, the Netherlands*

Abstract

Compact stellar systems such as Ultra-compact dwarfs (UCDs) and Globular Clusters (GCs) around galaxies are known to be the tracers of the merger events that have been forming these galaxies. Therefore, identifying such systems allows to study galaxies mass assembly, formation and evolution. However, in the lack of spectroscopic information detecting UCDs/GCs using imaging data is very uncertain. Here, we aim to train a machine learning model to separate these objects from the foreground stars and background galaxies using the multi-wavelength imaging data of the Fornax galaxy cluster in 6 filters, namely *u, g, r, i, J* and *Ks*. The classes of objects are highly imbalanced which is problematic for many automatic classification techniques. Hence, we employ Synthetic Minority Over-sampling to handle the imbalance of the training data. Then, we compare two classifiers, namely Localized Generalized Matrix Learning Vector Quantization (LGMLVQ) and Random Forest (RF). Both methods are able to identify UCDs/GCs with a precision and a recall of > 93% and provide relevances that reflect the importance of each feature dimension for the classification. Both methods detect angular sizes as important markers for this classification problem. While it is astronomical expectation that color indices of $u-i$ and $i-Ks$ are the most important colors, our analysis shows that colors such as $g-r$ are more informative, potentially because of higher signal-to-noise ratio. Besides the excellent performance the LGMLVQ method allows further interpretability by providing the feature importance for each individual class, class-wise representative samples and the possibility for non-linear visualization of the data as demonstrated in this contribution. We conclude that employing machine learning techniques to identify UCDs/GCs can lead to promising results. Especially transparent methods allow further investigation and analysis of importance of the measurements for the detection problem and provide tools for non-linear visualization of the data.

Keywords: galaxies: clusters: individual (Fornax), galaxies: star clusters, techniques: photometric, Machine Learning, explainable AI, ensemble learning

1. Introduction

Based on the current theories of galaxy formation and evolution, galaxies are formed hierarchically from the merger of low-mass galaxies that were formed earlier. In this picture, the dense stellar structures such as Ultra-compact dwarfs (UCDs) galaxies and Globular clusters (GCs), which are mostly found around galaxies and the core of galaxy clusters, are known to be the tracers of such merging events (Beasley, 2020). However, extragalactic UCDs and GCs around other galaxies than the Milky Way look like stars (point-sources), due to their distance and current limitations of observational equipment. Therefore identifying them through imaging is challenging (Jordán et al., 2009). To find these objects, it is necessary to measure their distances, which is only possible using spectroscopy and measuring their radial velocity. The Hubble-Lemaître Law says that more distant galaxies move faster away from us and thus astronomers measure the distance of galaxies by measuring their radial velocity. The latter can be measured using Doppler shifts of the absorption lines in the spectroscopic observations. Spectroscopy of astronomical objects, however, needs longer exposure times than imaging. In other words, spectroscopy for all the star-like objects (point-sources) is not feasible in practice (Voggel et al., 2020).

The recent advances in astronomical instrumentation and observations, without doubt, have provided us with a large amount of data to explore. Traditionally, the possible UCDs/GCs candidates are identified by multi-wavelength observations in a few optical filters (Cantiello et al., 2018). Once the candidates are found follow-up spectroscopy for selected nominees is carried out to measure the radial velocity and therefore the distance to confirm the identity of the objects (Pota et al., 2018). This implies that a more accurate UCD/GC selection makes the observation time shorter.

In the case of Fornax galaxy cluster, the second closest galaxy cluster to us, recent observations of optical and near-infrared have been available, which makes it possible to iden-

*Corresponding author
Email addresses: mohammadimathstar@gmail.com (Mohammad Mohammadi), kerstin.bunte@googlemail.com (Kerstin Bunte)
URL: www.cs.rug.nl/~kbunte (Kerstin Bunte)
[1]These authors contributed equally.

tify UCDs/GCs within the galaxy cluster. The optical part of the data has been used earlier to identify GCs in the cluster using various techniques (Cantiello et al., 2020; Prole et al., 2019; Angora et al., 2019). Recently, Muñoz et al., 2014 has shown that a combination of optical/near infrared filters improves the quality of identifying UCDs/GCs. However, this approach was not used until very recently, mostly because deep observations in the near-infrared are not as easy as the optical.

Due to the sheer amount of astronomical data automatic tools to analyse the data are highly desirable. Therefore, machine learning techniques get more and more attention among Astronomers recently, and they have been popularly explored for astronomical applications. The Random Forest (RF) (Carrasco et al., 2015; Gao et al., 2009) has been used to build a classifier for quasar identification and the Support Vector Machine (SVM) (Jones and Singal, 2017) has been employed to estimate the redshift. Other techniques used are k-nearest neighbor classifier (Li et al., 2008) for active galactic nuclei (AGN) detection, Support Vector Data Description (SVDD) (Mohammadi et al., 2019) for GC detection, and Artificial Neural Networks (ANN) (Ball et al., 2004; Barchi et al., 2020; Xiao et al., 2020), for galaxy morphological classification and AGN detection. Thus, a pletora of machine learning methods have been successfully applied in many astronomical applications and they encourage us to be extended to other cases.

A well-known family of prototype-based classifiers is Learning Vector Quantization (LVQ) that efficiently distinguish high dimensional multi-class problems. One of the main advantages of LVQ classifiers is the interpretability of the adaptive parameters. The learned prototypes, for example, can be investigated as typical representatives of the classes. While the original formulations of the LVQ family, such as *Generalized learning Vector Quantization* (GLVQ) (Sato and Yamada, 1995) use the Euclidean distance, more complex extensions, such as Generalized Matrix LVQ (GMLVQ) and the localized version LGMLVQ (Schneider et al., 2009b,a) employ adaptive distance metrics. The latter algorithms also allow further insights by visualization of the decision boundaries after training (Bunte et al., 2012; Bunte, 2011). In this contribution, we use LGMLVQ to classify three astronomical structures, namely foreground stars, UCDs/GCs and background galaxies, based on their optical (u, g, r, i) and near-infrared (J, Ks) measurements of the Fornax Deep Survey, VISTA Hemisphere Survey and ESO/VISTA archive. The method constructs non-linear decision boundaries and allows to evaluate the importance of features for each class individually. One major issue often faced in astronomical problems is the imbalance of the classes. The total number of known UCDs and GCs in the Fornax cluster identified in the data is just over 500, whereas the majority class contains about 5 times more instances. To tackle this challenge we apply oversampling techniques, such as Synthetic Minority Oversampling (SMOTE) (Chawla et al., 2002) and Borderline SMOTE (Han et al., 2005). In contrast to previous works we have both optical and near-infrared filters in the dataset. We use LGMLVQ to detect the classes of objects in potentially large amounts of high dimensional astronomical data and analyse the results. Moreover, we compare the performance using an ensemble of LGMLVQ classifiers.

The paper is organized as follows: In section 2 we describe the dataset, followed by the explanation of classifiers in section 3. Afterwards, in section 4 we describe the experiments and discuss the results. Finally we conclude in section 5 and provide inspirations for future work.

2. Astronomical Data and Preprocessing

The data used in this research is extracted from multi-wavelength wide astronomical surveys obtained from a combination of 6 filters, i.e optical filters (u, g ,r and i) and near-infrared filters (J and Ks). The optical u, g, r, i data was obtained from Fornax Deep Survey (FDS) using the ESO VLT survey telescope (VST), J from VISTA Hemisphere Survey (VHS, McMahon et al., 2013) using the VISTA telescope and Ks from the ESO/VISTA archival data. This data is described in detail in Saifollahi et al. (submitted to Monthly Notices of the Royal Astronomical Society). The data set provides photometric information in the direction of the Fornax galaxy cluster. After excluding larger objects, likely to be galaxies in the Fornax cluster or slightly more distant, any observed object in the dataset belong to one of the following 3 classes:

class 1: consists of 2826 background galaxies further away than the galaxies in the Fornax cluster,

class 2: denotes the class of interest consisting of 512 UCDs and GCs, and

class 3: contains 4399 nearby foreground stars located in our own galaxy, the Milky Way.

These classes are difficult to distinguish for several reasons: (1) the UCD and GC (class 2) observations are faint and ambiguous as they are engulfed between the two other classes, and (2) there is only a small number of confirmed examples of them.

The labelled data consists of 23 features. The coordinates of the objects in the sky are given in right ascension (RA) and declination (DEC) as a degree of point of observation from earth and illustrated in Fig. 1(c). Features $FWHM^*g$, $FWHM^*r$, $FWHM^*i$, $FWHM^*u$, $FWHM^*j$ and $FWHM^*k$, also known as the Full width half maximum, are the proxies for the angular sizes of the objects as seen from the corresponding filters. Followed by $u-g$, $u-r$, $u-i$, $u-J$, $u-Ks$, $g-r$, $g-i$, $g-J$, $g-Ks$, $r-i$, $r-J$, $r-Ks$, $i-J$, $i-Ks$ and $J-Ks$, which are the colour indices indicating the emission of the astronomical object in logarithmic scale, known as magnitude and correlated to physical properties like age and metallicity. In total, the data consists of 7737 complete observations that are used in the analysis.

Figure 1 shows the Eigenvalue profile of the data obtained using the unsupervised Principal Component Analysis (PCA) in panel (b). The two major Eigenvalues explain 88% of the data's variance and the corresponding two-dimensional projection is depicted in panel (a). The large areas of overlap and imbalance of the classes is clearly visible. Especially the latter states a problem for most classification techniques and thus re-sampling techniques, such as Synthetic Minority Over-sampling (SMOTE) (Chawla et al., 2002) and Borderline-SMOTE (Han et al., 2005), are investigated as preprocessing

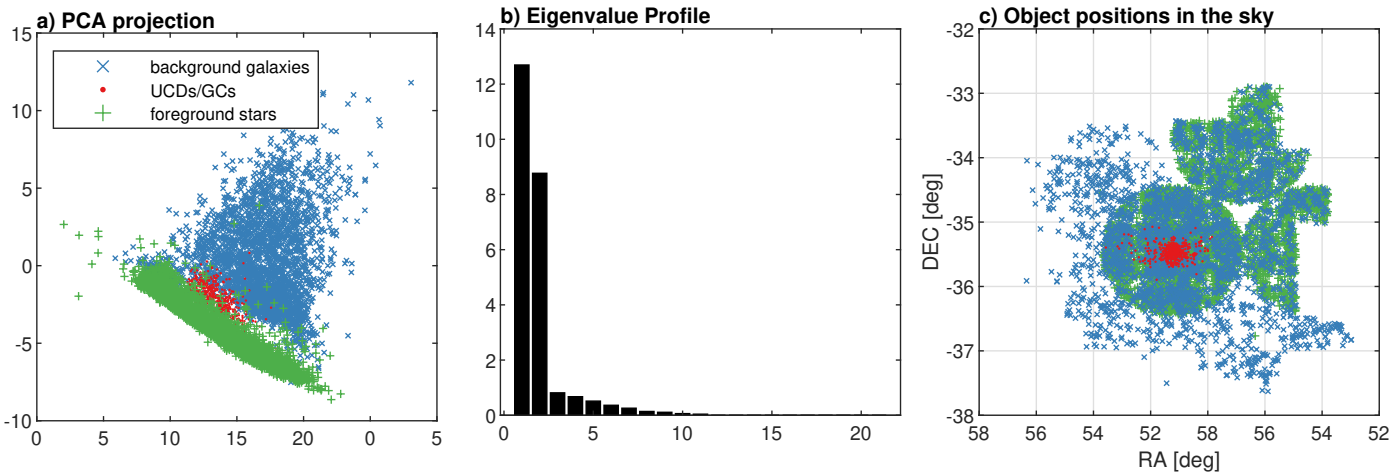

Figure 1: PCA-projection of the data colouring the background galaxies, UCDs/GCs and foreground stars (blue, red and green, classes 1-3), respectively (panel a). Corresponding Eigenvalue profile (panel b) and position of the objects in the sky in right ascension (RA) and declination (DEC) (panel c).

steps. SMOTE increases the population of the minority classes by generating synthetic samples as weighted convex combination between random samples and its nearest neighbours. Instead of choosing random samples Borderline-SMOTE specifically takes points near the boundaries between the classes which are more likely to be misclassified. In this contribution we investigate and compare both methods as preprocessing to balance the classes.

3. Methods

In this section, we introduce the state-of-the-art methodology used with focus on localized adaptive distance metrics in Learning Vector Quantization (LVQ) and corresponding interpretability. Moreover, we present a comparison between Random Forest (which is an ensemble of Decision Trees) and an ensemble of the Learning Vector Quantization models.

3.1. Learning Vector Quantization (LVQ)

We assume $\{(\boldsymbol{\xi}_i, y_i)\}_{i=1}^n$ denote the training set, where $\boldsymbol{\xi}_i \in \mathbb{R}^N$ and $y_i \in \{1, \cdots, C\}$ represent i-th data point and its class label, respectively. An LVQ classifier models the distribution of classes via a set of labelled prototypes $\{(\boldsymbol{\omega}^j, c(\boldsymbol{\omega}^j))\}_{j=1}^m$, where $c(\boldsymbol{\omega}^j)$ is the label of the respective prototype. These prototypes tessellate the data space into smaller regions, called Voronoi cells, enclosing data points for which the respective prototype is closer than any other. Classification follows a *nearest prototype scheme*, meaning any data point (including new ones) is assigned the class label of the nearest prototype. To find prototype positions that minimize the classification error E, *Generalized learning Vector Quantization* (GLVQ) (Sato and Yamada, 1995) introduced the following cost function, aiming at large margin optimization for better generalization:

$$E = \sum_{i=1}^{n} \Phi(\mu_i) \quad \text{with} \quad \mu_i = \frac{d(\boldsymbol{\xi}_i, \boldsymbol{\omega}^J) - d(\boldsymbol{\xi}_i, \boldsymbol{\omega}^K)}{d(\boldsymbol{\xi}_i, \boldsymbol{\omega}^J) + d(\boldsymbol{\xi}_i, \boldsymbol{\omega}^K)} \ , \qquad (1)$$

with Φ being a monotonically increasing function. We used the identity $\Phi(x) = x$ throughout this contribution. Furthermore, $d(\boldsymbol{\xi}_i, \boldsymbol{\omega}^J)$ denotes the distance of the data point $\boldsymbol{\xi}_i$ from the closest prototype with the same label $y_i = c(\boldsymbol{\omega}^J)$ and $d(\boldsymbol{\xi}_i, \boldsymbol{\omega}^K)$ the distance to the closest prototype with a different class label. The value of $\mu_i \in [-1, 1]$ shows the confidence of the classification. The cost function is non-convex and typically gradient techniques, such as stochastic gradient descent (Bunte, 2011; Schneider et al., 2009a) are utilized to minimize the costs E.

From Eq. (1) it is clear that the the distance measure d plays a major role for the performance of LVQ classifiers. While the Euclidean distance is a common choice all dimensions contribute equally in it, which has drawbacks in capturing underlying data semantics (Schneider et al., 2009a) in noisy high-dimensional and heterogeneous data spaces. As such it is not capable to reflect if features differ in importance for the classification task at hand. Therefore, Hammer and Villmann proposed to incorporate a weighting factor for each feature dimension that is adapted during training:

$$d^{\Lambda}(\boldsymbol{\xi}_i, \boldsymbol{\omega}^j) = (\boldsymbol{\xi}_i - \boldsymbol{\omega}^j)^{\top} \Lambda (\boldsymbol{\xi}_i - \boldsymbol{\omega}^j) \ , \qquad (2)$$

where the weight matrix Λ, also referred to as *relevance matrix*, is a diagonal matrix with 0 in the off-diagonals and λ_i on the diagonal with $\sum_i \lambda_i = 1$. These relevance weights indicate the discriminative contribution of each feature dimensions, which could facilitate decreasing influence or pruning of redundant, noisy or ambiguous feature dimensions. This concept can be further extended to more complicated metric tensors with adaptive off-diagonal elements, namely Generalized Matrix LVQ (GMLVQ) (Schneider et al., 2009b,a), Limited Rank Matrix LVQ (LiRaM LVQ) (Bunte et al., 2012; Bunte, 2011) and localized versions with prototype-wise or class-wise matrices called Localized GMLVQ (LGMLVQ) (Schneider et al., 2009b). All algorithms are made publicly available in Matlab and can be found at `https://github.com/kbunte/LVQ_toolbox`.

The overlapping class regions as shown in the PCA projection Figure 1(a) intuitively suggest non-linear class boundaries and hence the localized adaptive metrics are more suitable and

therefore the focus for this paper. Local metric tensors allow to learn localized dissimilarities with respect to the specific class prototypes using a local transformation matrix Ω^j thus defining the non-linear decision boundaries. The localized distance metric is defined as:

$$d^{\Lambda^j}(\boldsymbol{\xi}_i, \boldsymbol{\omega}^j) = (\boldsymbol{\xi}_i - \boldsymbol{\omega}^j)^\top \Lambda^j (\boldsymbol{\xi}_i - \boldsymbol{\omega}^j) \ , \tag{3}$$

where $\Lambda^j = \Omega^{j\top}\Omega^j$ using the adaptive matrix $\Omega^j \in \mathbb{R}^{M\times N}$ with $M \leq N$ to guarantee that Λ^j is positive semi-definite. The cost function therefore reads as follows:

$$E_{\mathrm{LGMLVQ}} = \sum_i^n \Phi(\mu^i_{\mathrm{local}}) \quad \text{where} \quad \mu^i_{\mathrm{local}} = \frac{d^{\Lambda^J} - d^{\Lambda^K}}{d^{\Lambda^J} + d^{\Lambda^K}} \tag{4}$$

where d^{Λ^J} and d^{Λ^K} are the distances of the point $\boldsymbol{\xi}_i$ from the closest correct and incorrect prototypes respectively. The update rules are described in detail in Schneider et al. (2009a); Bunte (2011). Besides allowing non-linear decision boundaries and therefore learning of more complex classification problems, the localized matrices furthermore enable the investigation of localized or class-wise relevances, marked on each diagonal of Λ^j, identifying features that are important for the classification of each class respectively (Schneider et al., 2009b).

3.2. Ensemble methods

With increasing complexity, classifiers get more powerful showing impressive performance in practice. However, at the same time they often show overfitting effects in which the performance on training data is near perfect but it decreases facing new data samples not seen before. This decreased *generalization* performance is often tackled using ensemble methods, which combine several classifiers to assign a class label to a new data instance to overcome the limitations of a single model. In order to see the effect of ensemble learning, we explore an ensemble of LGMLVQ models and compare it to Random Forest (RF). Random forest (Breiman, 2001) bags an ensemble of decision trees for classification that vote for the most popular class for a given input. The trees within, are constructed from the bootstrap examples that are sampled independently with replacement and random feature selection at the splits following a common distribution while growing to maximum depth with no pruning. The stochastic strategy is robust to outliers and improves the performance of random forest because of the law of large numbers from combination of rather low performing/weak decision tree classifiers (Breiman, 2001). In addition to classification, feature relevance values that are insightful can be extracted from a random forest for comparison with the LVQ method. For a fair comparison with RF the ensemble of LGMLVQ models is constructed from the same training bootstrap samples used in each decision tree in the random forest. This will result in as many LGMLVQ models as the number of decision trees for each cross validation fold. The results are then aggregated through majority voting (Ranawana and Palade, 2006).

3.3. Interpretability

For many applications it is crucial for machine learning models to be interpretable, such that the domain expert is able to examine the significance of the resulting trained model and its suitability for classification tasks. Intrinsic model interpretability can be understood as how understandable the internals of a model and its output are to users (Gilpin et al., 2018). It is furthermore suggestively explained by Backhaus and Seiffert (Backhaus and Seiffert, 2014) through a three-fold criteria of the model's 1) feature selection capability, 2) ability to define class representatives, such as prototypes and 3) encoding of classification boundary information. Interpretability for the random forest is achieved through random permutation of a feature's observations (Breiman, 2001; Strobl et al., 2007) for the out of bag samples and estimating the corresponding decision tree's accuracy with the permuted features. Here, more discriminative features are easily identified, since they have significant effect on the classification error. The out of bag predictor importance uncovers the individual impact of the features and the information could similarly be used for feature selection and understanding the random forest's classification. On the other hand, the adaptive LVQ methods satisfy all three criteria by: 1) Feature selection by means of the diagonal of the metric tensors, Λ from GMLVQ and the local Λ^j in LGMLVQ, that represents individual feature contribution that could be used as feature pruning criteria. 2) Prototype feature values used to classify novel observations are learned during the model training, which subsequently become typical representatives of their corresponding classes. 3) The decision boundaries for classification can be extracted and visualised, for example by linear projection of the data samples and the resulting class prototypes using Ω from GMLVQ. Nonlinear visualizations using the localized variants LiRaMLVQ and LGMLVQ are also possible and the latter is outlined in the following.

3.4. Nonlinear visualization with charting

Visualization can be useful to get a holistic view of the data and identify difficult instances. From the definition of $\Lambda^j = \Omega^{j\top}\Omega^j$ in Eq. (3), we see that the distance metric first transforms data points using the following local linear projections:

$$P_j : \boldsymbol{\xi} \rightarrow \Omega^{j\top}(\boldsymbol{\xi} - \boldsymbol{\omega}^j) \ .$$

For specific cases $M \in \{2, 3\}$, the projected data points can be visualized, which can be used for discriminant visualization of the data based on the space the classification takes place in. However, since the localized metric provides several projections for each sample, it is challenging to study the outputs directly. In order to tackle this problem, (Bunte et al., 2009) proposed a post-processing step which combines local projections using charting (Brand, 2003) to form a global nonlinear embedding of the data:

$$\boldsymbol{\xi} \rightarrow \sum_j p_j(\boldsymbol{\xi}) B_j(P_j(\boldsymbol{\xi})) \ .$$

Here, $B_j(.) : \mathbb{R}^M \rightarrow \mathbb{R}^M$ is an affine transformation and $p_j(\boldsymbol{\xi})$ is the responsibility of local transformation $\boldsymbol{\omega}^j$ for the data sample $\boldsymbol{\xi}$ with $\sum_j p_j(\boldsymbol{\xi}) = 1$. More details about how to compute prototypes' responsibilities and affine transformations can be found

in (Bunte et al., 2009)[2]. Using this nonlinear embedding we can easily project data to 2 (or 3) dimensions for further inverstigation of the overlapping regions and difficult samples.

4. Experiments and Discussion

This section shows the results and discussion of the experiments conducted with the localized adaptive distance metric LVQ method (LGMLVQ) coupled with presence or absence of resampling as a pre-processing step. The performance and feature importance is compared with Random Forest (RF). The corresponding feature relevances are examined in comparison with conventional astronomical expectations.

4.1. Experimental setup and Evaluation Measures

The experiments are set up within a 10-fold cross validation where the data observations are divided up into a 90/10 random but stratified training and test splits with each individual class preserving its sample frequency. For the distance based classifiers, such as LGMLVQ, each training set is normalised via Z-score transformation, e.g. zero mean and unit standard deviation, with the same parameters used to transform the respective test set. Decision trees and RF are not distance based and build instead rules on the features directly and therefore do not require transformative pre-processing in general. However, the RF is very sensitive to class imbalance, which should be handled before training. Therefore, resampling of the training data can be introduced to reduce or eliminate the imbalance of the classes. In this contribution we compare two strategies, namely the Synthetic Minority Oversampling Technique (SMOTE) and Borderline-SMOTE, creating new feature vectors using the training samples of each minority class until their amount matches the size of the majority class. The created synthetic minority samples lead to balanced classes to be used for training of the classification models.

The different models have different hyper-parameters. In the experiments we train the RF with 100 decision trees, sampling with replacement of 0.75 fractions of the training set and using the bagging aggregation method (Breiman, 2001). The LVQ models provide several hyper-parameters to control the methods complexity, such as the number of prototypes, number of metric tensors and their rank determining the projection dimension saving memory and enabling visualization. Due to the non-linearity of the problem we use the localized most powerful version of the LVQ family, namely LGMLVQ (Schneider et al., 2010). with one prototype per class and regularization of 10^{-5}. In order to choose the lower dimensional projection dimension we train the method using full metric tensors constructed using $\Omega^j \in \mathbb{R}^{M\times N}$ with $M = N$. Subsequently, an Eigenvalue decomposition of the trained $\Lambda^j = U^j\Sigma^j U^{j(-1)}$ with diagonal matrices $\Sigma^j_{ii} = \sigma^j_i$ providing information about the intrinsic dimensionality of the classification problem by counting the Eigenvalues $M = \max_j \sum_{i=1}^{N}[\sigma^j_i > \epsilon]$ significantly larger than 0. The model is than retrained with the rank M reduced to maximal number obtained from the matrices. The RF is an ensemble of decision trees and therefore we also build an ensemble of the trained LGMLVQ models from the same bootstrap samples used the 100 decision trees for direct comparison and the results are aggregated through majority voting. Resulting models were averaged across cross-validation runs and evaluated with focus on the class of interest.

[2]The code is available at `https://github.com/kbunte/LVQ_toolbox`

Table 1: Three-class confusion matrix and formulae to obtain the False Negatives, True Positives, False Positives and True Negatives (FP_b, TP_b, FP_b and TN_b) to calculate Precision, Specificity and Sensitivity with respect to class b.

Actual \ Predicted	Class 1	Class 2	Class 3
Class 1	C_{11}	C_{12}	C_{13}
Class 2	C_{21}	C_{22}	C_{23}
Class 3	C_{31}	C_{32}	C_{33}

$$FN_b = \sum_{\substack{f=1\\f\neq b}}^{3} C_{bf} \qquad FP_b = \sum_{\substack{f=1\\f\neq b}}^{3} C_{fb}$$
$$TN_b = \sum_{\substack{f=1\\f\neq b}}^{3}\sum_{\substack{q=1\\q\neq b}}^{3} C_{fqx} \qquad TP_b = C_{bb} \tag{5}$$

Table 2: Macro average performance (standard deviation) for techniques LGMVLQ (T=L) and Random Forest (T=RF) (abbreviated by ${}_{\{\varnothing|B|R\}}\mathrm{T}_{M}^{\{\varnothing|\mathrm{Ens}\}}$).

	Prec.	Spec.	Sens.	Train accur	Test accur
L_2	0.969 (0.012)	0.986 (0.004)	0.930 (0.021)	0.985 (0.001)	0.984 (0.004)
${}_{\mathrm{R}}\mathrm{L}_2$	0.935 (0.020)	0.987 (0.005)	0.963 (0.020)	0.983 (0.001)	0.982 (0.005)
${}_{\mathrm{B}}\mathrm{L}_2$	0.889 (0.026)	0.979 (0.007)	0.950 (0.027)	0.971 (0.005)	0.971 (0.006)
${}_{\mathrm{R}}\mathrm{L}_2^{\mathrm{E}}$	0.937 (0.019)	0.987 (0.005)	0.962 (0.020)	0.983 (0.001)	0.983 (0.005)
${}_{\mathrm{R}}\mathrm{RF}^{\mathrm{E}}$	0.950 (0.018)	0.989 (0.005)	0.964 (0.018)	0.999 (0.000)	0.986 (0.005)

For evaluation output statistics are generated after prediction with the models, i.e. training and test errors, and their standard deviations. Since this is a multi-class classification problem and the major interest is in Class 2 UCD/GC objects, we extract the evaluation measures for each class and build the macro averaged accuracies to evaluate the performance across the different classes and eliminate bias of majority populated class. We report also binary class measures, such as precision, sensitivity and specificity, for the class of interest versus all others classes combined to represent the algorithms performance in classifying the novel test data. The confusion matrix as illustrated in Table 1, can be used to evaluate the classification performance providing detailed information about the accuracy for each class and the nature of misclassifications. From it one can extract the binary measures, namely false positives (FP), true positives (TP), false negatives (FN) and true negatives (TN) as shown below Table 1 in Eq. (5). Additionally, the false positive rate and true positive rate of the test set are plotted on a Receiver Operating Characteristic (ROC) curve (Fawcett, 2006). This curve shows the model's discriminative ability and the Area Under the Curve (AUC) summarizes its overall performance for the classification of the class of interest.

4.2. Discusssion

In this section we summarize the results of the methods on the astronomical classification problem. The hyperparameter

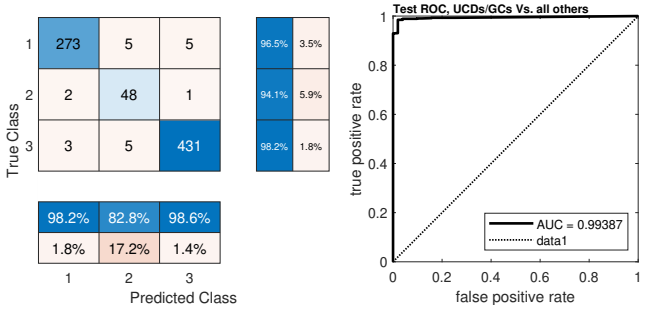

Figure 2: Panel (a) shows the test confusion matrix for the ${}_{\mathrm{R}}\mathrm{L}_2$ model and panel (b) shows the corresponding ROC curve of minority class 2 vs. all the other classes and the corresponding AUC value of 0.99387.

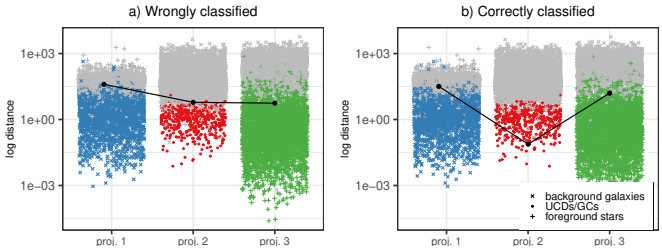

Figure 3: Wrong (a) and correct (b) classification of two minority class observations. Each shows the distances of all samples projected using the local Ω^j of each prototype, highlighting the samples with corresponding label j in colour.

settings for the techniques LGMLVQ (T=L) and Random Forest (T=RF) is abbreviated by ${}_{\{\varnothing|B|R\}}\mathrm{T}_M^{\{\varnothing|\mathrm{E}\}}$. Here the preprocessing is marked by letters R and B for resampling with SMOTE or Borderline-SMOTE, the subscript M denotes the rank for the local metric tensors and the superscript E denotes the results of an ensemble of 1000 models is reported. The performance of the LGMLVQ models is excellent already with an intrinsic dimensionality of $M = 2$ even without resampling, evident from the sensitivity and precision for the minority class, as shown in Table 2. Resampling improves the correct classification by 14.79% for the minority class as demonstrated by an increase in the sensitivity. However, precision reduces indicating that there are more false positives brought about by SMOTE resampling, which is an acceptable trade-off. Figure 2 panel a) and b) show the test performance of ${}_{\mathrm{R}}\mathrm{L}_2$ with only 3 false negatives for the minority class 2 of UCDs and GCs. Contrary to our expectations Borderline-SMOTE resampling does not perform better. This could be caused by the overlap of the classes which increases the difficulty to define a clear boundary and hence boundary resampling becoming ineffective. In summary the Random Forest and LVQ models show comparable performance. However, especially the latter is less complex and provides further insights into the results of the classification which will be discussed in the following.

As mentioned before the LVQ models are intrinsically interpretable and transparent in many regards. We can for example interpret the certainty of the classification by investigating the distance of each sample to all prototypes. To demonstrate this we project all samples and all prototypes with the local transformations Ω^j and compute the distance to each prototype

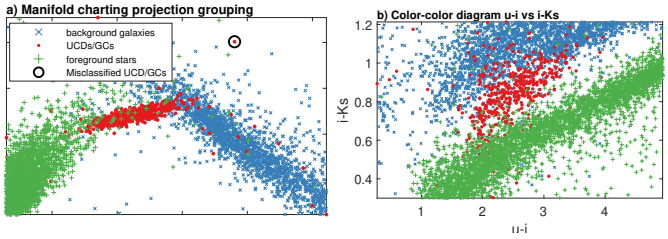

Figure 4: a) Manifold charting projection of the data by LGMLVQ (${}_{\mathrm{R}}\mathrm{L}_2$) and b) conventional color-color diagram used in photometric selection by Astronomy.

in the transformed space. Figure 3 visualizes these distances highlighting the samples with the same respective class label j in colour and of different classes in grey. We furthermore highlight in black the distances of an observations consistently misclassified in repeated training runs (panel a) and a correctly classified sample (panel b). The wrongly classified minority sample in (a) is within the boundary region where the classes 2 and 3 overlap as indicated by very similar distances to the prototypes of these classes. Panel (b), on the other hand, shows a typical example of a very certain correct classification where the sample is much closer to prototype 2 compared to the others. This transparently informs the user how sure the LGMLVQ classifier is with the assignment of a class label for each observation, which may also indicate samples with potentially controversial current label identification given the data at hand, marking them as candidates for further investigation.

Moreover, the local linear projections $\Omega^j \in \mathbb{R}^{M\times N}$ can be used for nonlinear visualization for $M \in \{2, 3\}$ using manifold charting as outlined in section 3.4. Hence we report very good performance for LGMLVQ using the rank $M = 2$, we show the resulting projected data of ${}_{\mathrm{R}}\mathrm{L}_2$ to complement the data analysis in Figure 4 panel a. The more distinctive separation provided by the LGMLVQ model, especially for the minority class, explains the efficiency and nuance of the method's classification performance as compared to the traditional astronomical color-color diagram as shown in panel b. This visualization shows the difficult regions and can be used to identify difficult observations, such as the circled class 2 sample in panel a, in the now reduced overlapping areas.

As mentioned before the RF and LGMLVQ models allow to extract the importance of the measurements for the classification problem. However, the Random Forest method is expensive in terms of memory costs and shows a clear tendency to overfit as seen in the test error being higher than the training error. Panel (b) in Figure 5 shows the dominance of the angular size features $FWHM^*u$, $FWHM^*g$, $FWHM^*r$, $FWHM^*i$, $FWHM^*J$ and $FWHM^*Ks$ in importance for the Random Forest classifier. In contrast to RF the LGMLVQ classifier extracts the feature relevances for each prototype and hence we can discuss also the relevance of the measurements for the classification of each class of objects in our astronomical data. Figure 5(a) shows the class-wise normalized feature importance profiles sorted by value of contribution with varying top relevant features further explaining the non-linearity and motivation for choice of local metric tensors. The top relevant fea-

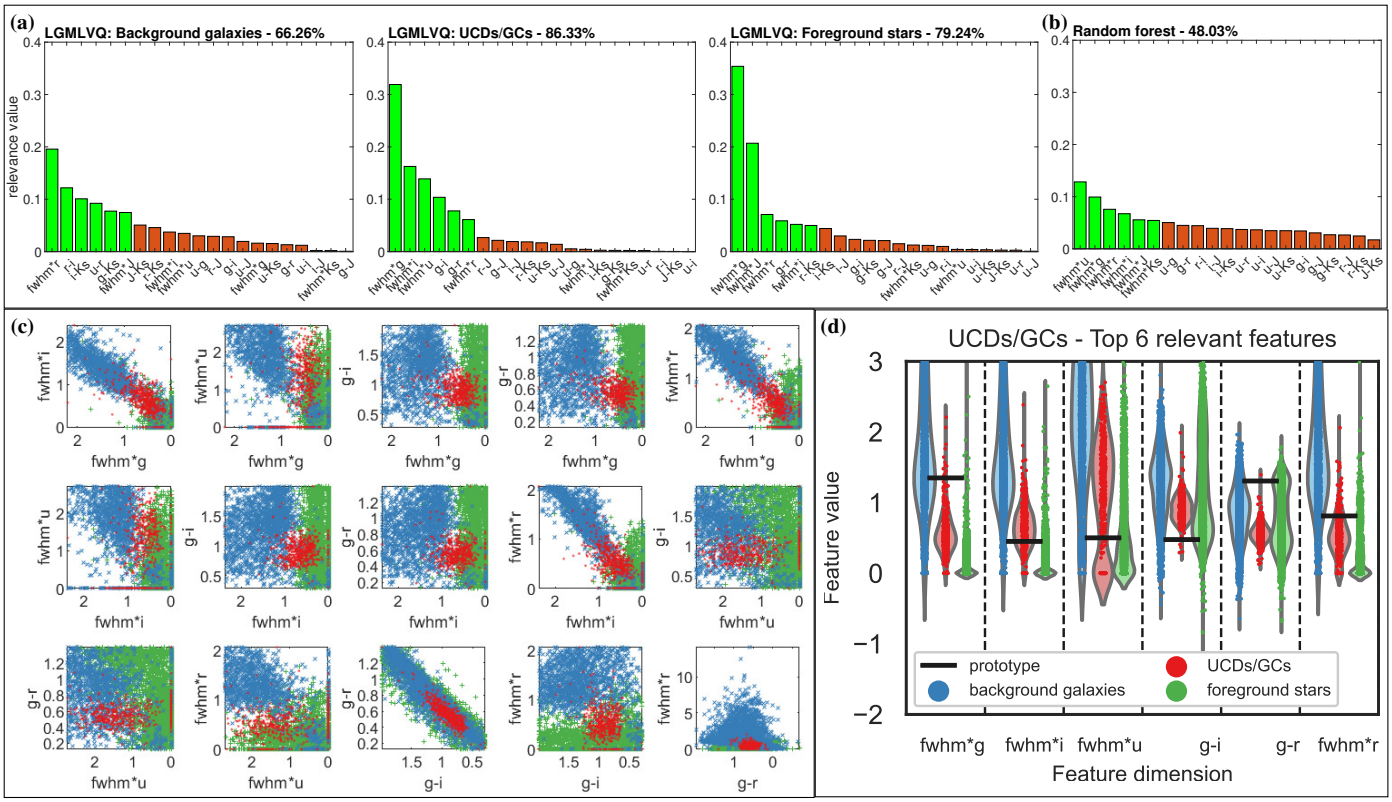

Figure 5: Panel (a): class-wise feature relevance profiles of LGMLVQ ${}_{\mathrm{R}}\mathrm{L}_2$ marking the top 6 in green and their corresponding percentage of contribution to the respective class and (b) the Random Forest (RF) feature relevance profile. Panel (c): pairwise plots of the ${}_{\mathrm{R}}\mathrm{L}_2$ top 6 features for the UCDs/GCs class 2 (red markers) with focus on the area covered by that class, and (d) corresponding violin plots with prototype positions relative to the feature value distribution for minority class 2.

tures for classifying the minority class of UCDs and GCs from this experiment dominantly consist of the angular size parameters, namely $FWHM^*g$, $FWHM^*i$, $FWHM^*u$ and $FWHM^*r$ and their pairwise correlation plots are visualized in panel (c) in Figure 5. Similarly panel (d) provides additional information in form of a violin plot for the 6 features most important for the classification of class 2, showing the distributions of the measurements of each class together with the value the class prototype exhibits after training. These angular size features are important for separating the majority of objects in class 1, whose objects are larger, from class 2 and 3 that have small sizes. The astronomical expectation is that the angular size cannot discriminate class 2 and 3 as illustrated by the model. However, in contrary to astronomical expectation, the angular sizes are found to be important to distinguish class 2 and 3 by both the LGMLVQ and RF, as shown in Figure 5. The disparity can be attributed to the measurement biases: The minority class 2 objects are faint (and lower in signal-to-noise ratio) and hence angular size measurements happen to be larger than the actual size which could introduce a bias to the data. Therefore, their angular size proxies, i.e. $FWHM^*g$, $FWHM^*i$, $FWHM^*u$ and $FWHM^*r$, could possess more discriminating information than colour indices.

Based on (Munoz et al., 2013), in the combined optical/near-infrared observations, the color indices of $u-i$ and $i-Ks$ are expected to be the most important features. In a simple view, the $u-i$ color is more sensitive to the age of an object while $i-Ks$ represents the metallicity (the amount of metals heavier than Helium) of the object. Other color indices, such as $g-i$, $g-r$, $r-J$ etc. could also partially carry these information but degenerate. In contrary, the observations in the u and Ks are harder to be carried out and often have a lower signal-to-noise ratio compared to the other filters. This makes the expected feature importance of $u-i$ and $i-Ks$ relatively uncertain. Values of $g-r$ and $g-i$ on the other hand are notably accurate.

To examine further the difficult misclassified UCD/GCs example circled in Figure 4 a), we visualize in Figure 6 the feature values (bottom panel) in comparison with the closest correct prototype (middle) and the corresponding local relevance profile (top panel). The 6 most important measurements for the classification are marked in green as before in each panel. This difficult example is quite different from the correct classifying prototype in the four most important feature dimensions $FWHM^*g$, $FWHM^*j$, $FWHM^*r$ and $g-r$, where the values are much smaller, which explains the misclassification. Such examples require more accurate measurements sizes (using deeper observations) to investigate whether the object indeed belongs to the expected class.

5. Conclusion

In this paper, we explore and compare two interpretable machine learning techniques, namely Localized General Matrix LVQ (LGMLVQ) and Random Forest (RF), for the analysis

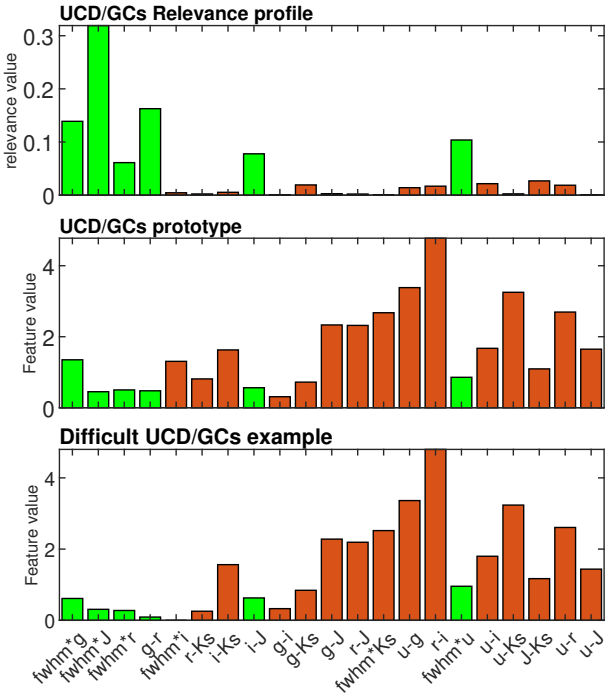

Figure 6: Difficult UCD/GCs example (bottom) in contrast with the the closest correct prototype (middle) and corresponding relevance profile (top). The 6 most relevant features for UCD/GCs LGMLVQ classification are marked green.

and classification of foreground stars and background galaxies versus UCDs and GCs. Due to the class of interest being highly underrepresented compared the former objects we also investigate the influence of Synthetic Minority Oversampling TEchnique (SMOTE) and its borderline extension on the classification performance. Localized distances allowing nonlinear decision boundaries within the data improves the classification in LGMLVQ, even in this situation where the classes largely overlap. LGMLVQ is also highly interpretable in the form of prototype class representatives and feature relevances which attach values to the contribution of a feature to classification. Additionally, the experiments uncover classification patterns in terms of feature relevances, which serve as discriminative markers for the classification.

The $u-i$ and $i-Ks$ colors are expected to be the most relevant colors for classification since they carry physical information on ages and metallities of astronomical objects. However, higher signal-to-noise ratio colors such as $g-i$ and $r-J$ in LGMLVQ and $u-g$ and $g-r$ in Random Forest are found to be more important for the data-driven classification. The importance of other colors compared to $u-i$ and $i-Ks$, that have almost 0 relevance contribution, implies that this disparity may be attributed to astronomically expected features having uncertain measurements, but also the correlation across the features, meaning that they partially contain the same critical information. Furthermore, angular size features $FWHM^*g$, $FWHM*i$, $FWHM^*u$ are identified by both methods independently as the most important features for the classification. We discuss that this outcome is due to a measurements biases of the faint sources of class 2 (UCDs/GCs).

In this work we showed that existing machine learning techniques can be used to identify UCDs/GCs in big astronomical data. These methods can handle the imbalance in the data and classify sources with a good performance. Subsequent analysis of the transparent techniques allows further insight and can provide valuable information for the astronomical experts to inform about possible biases in the data set. A future deeper data set with more accurate size and color measurements will most likely increase the performance of the automated classification even further.

Acknowledgments

This project has received financial support from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 721463 to the SUNDIAL ITN network.

References

Angora, G., Brescia, M., Cavuoti, S., Paolillo, M., Longo, G., Cantiello, M., Capaccioli, M., D'Abrusco, R., D'Ago, G., Hilker, M., Iodice, E., Mieske, S., Napolitano, N., Peletier, R., Pota, V., Puzia, T., Riccio, G., Spavone, M., 2019. Astroinformatics-based search for globular clusters in the Fornax Deep Survey. MNRAS 490, 4080–4106. doi:10.1093/mnras/stz2801, arXiv:1910.01884.

Backhaus, A., Seiffert, U., 2014. Classification in high-dimensional spectral data: Accuracy vs. interpretability vs. model size. Neurocomputing 131, 15–22.

Ball, N.M., Loveday, J., Fukugita, M., Nakamura, O., Okamura, S., Brinkmann, J., Brunner, R.J., 2004. Galaxy types in the sloan digital sky survey using supervised artificial neural networks. Monthly Notices of the Royal Astronomical Society 348, 1038–1046.

Barchi, P., de Carvalho, R., Rosa, R., Sautter, R., Soares-Santos, M., Marques, B., Clua, E., Gonçalves, T., de Sá-Freitas, C., Moura, T., 2020. Machine and deep learning applied to galaxy morphology-a comparative study. Astronomy and Computing 30, 100334.

Beasley, M.A., 2020. Globular Cluster Systems and Galaxy Formation. pp. 245–277. doi:10.1007/978-3-030-38509-5_9.

Brand, M., 2003. Charting a manifold, in: Advances in neural information processing systems, pp. 985–992.

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Bunte, K., 2011. Adaptive dissimilarity measures, dimension reduction and visualization. Ph.D. thesis.

Bunte, K., Hammer, B., Schneider, P., Biehl, M., 2009. Nonlinear discriminative data visualization., in: ESANN, pp. 65–70.

Bunte, K., Schneider, P., Hammer, B., Schleif, F.M., Villmann, T., Biehl, M., 2012. Limited rank matrix learning, discriminative dimension reduction and visualization. Neural Networks 26, 159 – 173. URL: http://www.sciencedirect.com/science/article/pii/S0893608011002632.

Cantiello, M., Grado, A., Rejkuba, M., Arnaboldi, M., Capaccioli, M., Greggio, L., Iodice, E., Limatola, L., 2018. A VST and VISTA study of globular clusters in NGC 253. A&A 611, A21. doi:10.1051/0004-6361/201731325, arXiv:1711.00805.

Cantiello, M., Venhola, A., Grado, A., Paolillo, M., D'Abrusco, R., Raimondo, G., Quintini, M., Hilker, M., Mieske, S., Tortora, C., Spavone, M., Capaccioli, M., Iodice, E., Peletier, R., Falcon Barroso, J., Limatola, L., Napolitano, N., Schipani, P., van de Ven, G., Gentile, F., Covone, G., 2020. The Fornax Deep Survey with VST. X. The catalog of sources in the FDS area, with an example study for globular clusters and background galaxies. arXiv e-prints , arXiv:2005.12085arXiv:2005.12085.

Carrasco, D., Barrientos, L., Pichara, K., Anguita, T., Murphy, D.N., Gilbank, D.G., Gladders, M.D., Yee, H.K., Hsieh, B.C., López, S., 2015. Photometric classification of quasars from rcs-2 using random forest. Astronomy & Astrophysics 584, A44.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. Smote: Synthetic minority over-sampling technique. J. Artif. Int. Res. 16, 321–357.

Fawcett, T., 2006. An introduction to roc analysis. Pattern recognition letters 27, 861–874.

Gao, D., Zhang, Y.X., Zhao, Y.H., 2009. Random forest algorithm for classification of multiwavelength data. Research in Astronomy and Astrophysics 9, 220.

Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L., 2018. Explaining explanations: An approach to evaluating interpretability of machine learning. CoRR URL: http://arxiv.org/abs/1806.00069.

Hammer, B., Villmann, T., 2002. Generalized relevance learning vector quantization. Neural Networks 15, 1059–1068.

Han, H., Wang, W.Y., Mao, B.H., 2005. Borderline-smote: A new over-sampling method in imbalanced data sets learning, in: Huang, D.S., Zhang, X.P., Huang, G.B. (Eds.), Advances in Intelligent Computing, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 878–887.

Jones, E., Singal, J., 2017. Analysis of a custom support vector machine for photometric redshift estimation and the inclusion of galaxy shape information. Astronomy & Astrophysics 600, A113. URL: http://dx.doi.org/10.1051/0004-6361/201629558.

Jordán, A., Peng, E.W., Blakeslee, J.P., Côté, P., Eyheramendy, S., Ferrarese, L., Mei, S., Tonry, J.L., West, M.J., 2009. The ACS Virgo Cluster Survey XVI. Selection Procedure and Catalogs of Globular Cluster Candidates. The Astrophysical Journal Supplement Series 180, 54–66. doi:10.1088/0067-0049/180/1/54.

Li, L., Zhang, Y., Zhao, Y., 2008. k-nearest neighbors for automated classification of celestial objects. Science in China Series G: Physics, Mechanics and Astronomy 51, 916–922.

McMahon, R.G., Banerji, M., Gonzalez, E., Koposov, S.E., Bejar, V.J., Lodieu, N., Rebolo, R., VHS Collaboration, 2013. First Scientific Results from the VISTA Hemisphere Survey (VHS). The Messenger 154, 35–37.

Mohammadi, M., Petkov, N., Bunte, K., Peletier, R.F., Schleif, F.M., 2019. Globular cluster detection in the gaia survey. Neurocomputing 342, 164–171.

Muñoz, R.P., Puzia, T.H., Lançon, A., Peng, E.W., Côté, P., Ferrarese, L., Blakeslee, J.P., Mei, S., Cuillandre, J.C., Hudelot, P., Courteau, S., Duc, P.A., Balogh, M.L., Boselli, A., Bournaud, F., Carlberg, R.G., Chapman, S.C., Durrell, P., Eigenthaler, P., Emsellem, E., Gavazzi, G., Gwyn, S., Huertas-Company, M., Ilbert, O., Jordán, A., Läsker, R., Licitra, R., Liu, C., MacArthur, L., McConnachie, A., McCracken, H.J., Mellier, Y., Peng, C.Y., Raichoor, A., Taylor, M.A., Tonry, J.L., Tully, R.B., Zhang, H., 2014. The Next Generation Virgo Cluster Survey-Infrared (NGVS-IR). I. A New Near-Ultraviolet, Optical, and Near-Infrared Globular Cluster Selection Tool. The Astrophysical Journal Supplement Series 210, 4. doi:10.1088/0067-0049/210/1/4, arXiv:1311.0873.

Munoz, R.P., Puzia, T.H., Lançon, A., Peng, E.W., Cote, P., Ferrarese, L., Blakeslee, J.P., Mei, S., Cuillandre, J.C., Hudelot, P., et al., 2013. The next generation virgo cluster survey-infrared (ngvs-ir). i. a new near-ultraviolet, optical, and near-infrared globular cluster selection tool. The Astrophysical Journal Supplement Series 210, 4.

Pota, V., Napolitano, N.R., Hilker, M., Spavone, M., Schulz, C., Cantiello, M., Tortora, C., Iodice, E., Paolillo, M., D'Abrusco, R., Capaccioli, M., Puzia, T., Peletier, R.F., Romanowsky, A.J., van de Ven, G., Spiniello, C., Norris, M., Lisker, T., Munoz, R., Schipani, P., Eigenthaler, P., Taylor, M.A., Sánchez-Janssen, R., Ordenes-Briceño, Y., 2018. The Fornax Cluster VLT Spectroscopic Survey - I. VIMOS spectroscopy of compact stellar systems in the Fornax core region. MNRAS 481, 1744–1756. doi:10.1093/mnras/sty2149, arXiv:1803.03275.

Prole, D.J., Hilker, M., van der Burg, R.F.J., Cantiello, M., Venhola, A., Iodice, E., van de Ven, G., Wittmann, C., Peletier, R.F., Mieske, S., Capaccioli, M., Napolitano, N.R., Paolillo, M., Spavone, M., Valentijn, E., 2019. Halo mass estimates from the globular cluster populations of 175 low surface brightness galaxies in the Fornax cluster. MNRAS 484, 4865–4880. doi:10.1093/mnras/stz326, arXiv:1901.09648.

Ranawana, R., Palade, V., 2006. Multi-classifier systems: Review and a roadmap for developers. Int. J. Hybrid Intell. Syst. 3, 35–61.

Sato, A., Yamada, K., 1995. Generalized learning vector quantization, in: Proceedings of the 8th International Conference on Neural Information Processing Systems, MIT Press, Cambridge, MA, USA. p. 423–429.

Schneider, P., Biehl, M., Hammer, B., 2009a. Adaptive relevance matrices in learning vector quantization. Neural Comput. 21, 3532–3561. URL: https://doi.org/10.1162/neco.2009.11-08-908.

Schneider, P., Biehl, M., Hammer, B., 2009b. Distance learning in discriminative vector quantization. Neural Computation 21, 2942–2969. URL: https://doi.org/10.1162/neco.2009.10-08-892.

Schneider, P., Bunte, K., Stiekema, H., Hammer, B., Villmann, T., Biehl, M., 2010. Regularization in matrix relevance learning. IEEE Transactions on Neural Networks 21, 831–840.

Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics 8, 25.

Voggel, K.T., Seth, A.C., Sand, D.J., Hughes, A., Strader, J., Crnojevic, D., Caldwell, N., 2020. A Gaia-based Catalog of Candidate Stripped Nuclei and Luminous Globular Clusters in the Halo of Centaurus A. The Astrophysical Journal Supplement Series 899, 140. doi:10.3847/1538-4357/ab6f69, arXiv:2001.02243.

Xiao, H., Cao, H., Fan, J., Costantin, D., Luo, G., Pei, Z., 2020. Efficient fermi source identification with machine learning methods. Astronomy and Computing , 100387.