Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **457,** 3119–3132 (2016) doi:10.1093/mnras/stw157

An analysis of feature relevance in the classification of astronomical transients with machine learning methods

A. D'Isanto,[1,2⋆] S. Cavuoti,[3⋆] M. Brescia,[3⋆] C. Donalek,[4] G. Longo,[1] G. Riccio[3] and S. G. Djorgovski[4,5]

[1] *Department of Physical Sciences, University of Napoli Federico II, via Cinthia 9, I-80126 Napoli, Italy*
[2] *Heidelberg Institute for Theoretical Studies (HITS), Schloss-Wolfsbrunnenweg 35, D-69118 Heidelberg, Germany*
[3] *INAF – Astronomical Observatory of Capodimonte, via Moiariello 16, I-80131 Napoli, Italy*
[4] *Center for Data Driven Discovery, California Institute of Technology, 1200 E. California Blvd, 91125 Pasadena, USA*
[5] *Department of Astronomy, California Institute of Technology, 1216 East California bvd, Pasadena, CA 91125, USA*

Accepted 2016 January 15. Received 2016 January 15; in original form 2015 October 30

ABSTRACT

The exploitation of present and future synoptic (multiband and multi-epoch) surveys requires an extensive use of automatic methods for data processing and data interpretation. In this work, using data extracted from the Catalina Real Time Transient Survey (CRTS), we investigate the classification performance of some well tested methods: Random Forest, MultiLayer Perceptron with Quasi Newton Algorithm and K-Nearest Neighbours, paying special attention to the feature selection phase. In order to do so, several classification experiments were performed. Namely: identification of cataclysmic variables, separation between galactic and extragalactic objects and identification of supernovae.

Key words: methods: data analysis – novae, cataclysmic variables – supernovae: general – stars: variables: general – stars: variables: RR Lyrae.

1 INTRODUCTION

The advent of a new generation of multi-epoch and multiband (synoptic) surveys has opened a new era in astronomy allowing us to study with unprecedented accuracy the physical properties of variable sources. The potential of these new digital surveys, both in terms of new discoveries as well as of a better understanding of already known phenomena, is huge. For instance, the Catalina Real-Time Transient Survey (CRTS, Drake et al. 2009) in less than eight years of operation, enabled the discovery of $\sim$2400 supernova (SN), $\sim$1200 cataclysmic variables (CV), $\sim$2800 active galactic nuclei (AGN), as well as to identify brand new phenomena such as binary black holes (Graham et al. 2015) and peculiar types of SN (Drake et al. 2010). A discovery trend which is expected to continue and even increase when new observing facilities such as the Large Synoptic Telecope (Closson Ferguson 2015), and the Square Kilometre array (Yahya et al. 2015) become operational.

With these new instruments, however, both size of the data and event discovery rates are expected to increase, from the current $\sim$10–10^2 events per night, up to $\sim 10^5$–10^7. Only a small fraction of these events will be targeted by dedicated follow-ups and therefore it will become crucial to disentangle potentially interesting events from lesser ones. With data volumes already in the terabyte and petabyte domain, the discrimination of time-critical information has already exceeded the capabilities of human operators and also crowds of citizen scientists cannot match the task. A viable approach is therefore to automatize each step of the data acquisition, processing and understanding tasks. In this work, with 'data understanding', we mean the identification of transients and their classification into broad classes, such as periodic versus non-periodic, SN, CV stars, etc.

Many efforts have been made to apply a variety of machine learning (ML) methods to classification problems (du Buisson et al. 2015; Rebbapragada 2014; Goldstein et al. 2015; Wright et al. 2015).

Real time analysis can be performed using different methods, among which we shall just recall those based on Random Forest (RF; Breiman 2001) and on Hierarchical Classification (Lo et al. 2013).

Off-line classification, being less critical in terms of computing time, can be performed with many different types of classifiers. It is common practice to distinguish between supervised and unsupervised methods, depending on whether a previously classified sample is or is not used for the training phase. In the supervised category, we have, for instance, Bayesian Network (Castill, Gutierrez & Hadi 1997), Support Vector Machines (Chang & Lin 2011), K-Nearest Neighbours (KNN; Hastie, Tibshirani & Friedman 2001), RF (Breiman 2001), and Neural Networks (McCulloch & Pitts 1943). While in the unsupervised family, we mention Gaussian Mixture Modelling (McLachlan & Peel 2001), and Self-Organizing Maps (Kohonen 2001).

[⋆] E-mail: antonio.Disanto@h-its.org (ADI); stefano.cavuoti@gmail.com (SC); brescia@oacn.inaf.it (MB)

Published by Oxford University Press on behalf of the Royal Astronomical Society

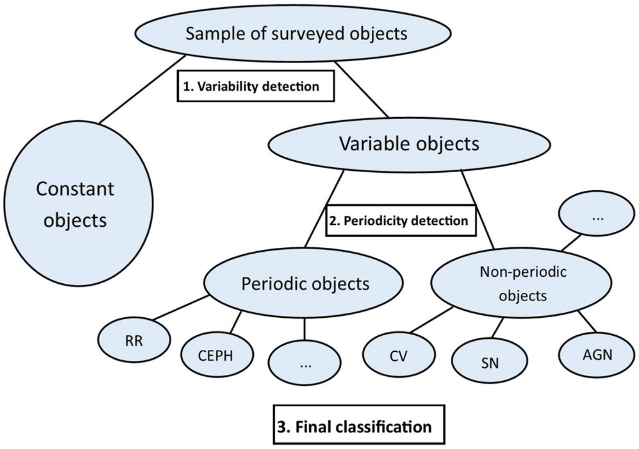


Figure 1. An adapted version of the scheme presented in Dubath (2012) for a general classification of variable objects.

In this work, we shall focus on off-line classification, making use of three different machine learning methods, namely: the MultiLayer Perceptron with Quasi-Newton Algorithm (MLPQNA, Brescia et al. 2012), the RF (Breiman 2001) and the KNN (Hastie et al. 2001). Most of the presented work was performed in the framework of the Data Mining & Exploration Web Application REsource (DAMEWARE; Brescia et al. 2014) infrastructure and the PhotoRaptor public tool (Cavuoti et al. 2015).

The paper is structured as it follows: in Section 2, we present the data and introduce the features extracted for the analysis. In Section 3, we briefly describe the ML methods used for the experiments detailed in Section 4. Results are discussed in 5.

2 THE DATA

In what follows, we shall divide objects according to a simplified version (see Fig. 1) of the semantic tree described in Eyer & Mowlavi (2008). From this scheme, it emerges quite naturally, the need to split the classification task in at least three steps (e.g. Dubath 2012). In the first step, variable objects (the transients) are disentangled from normal, non-variable stars. In the second step, periodic objects are separated from non-periodic objects and, finally, in the third and last step, one can proceed to the final classification of the objects.

In this work we make use of 1619 light curves extracted from the CRTS (Drake et al. 2009) public archive. CRTS is a synoptic astronomical survey that repeatedly covers 33 000 deg^2 of the sky with the main goal of discovering rare and interesting transient phenomena. The survey utilizes data taken in only one band (V) by the three dedicated telescopes of the highly successful Catalina Sky Survey near-Earth objects (NEO) project and detects and openly publishes all transients within minutes of observation so that all astronomers may follow ongoing events.

The sample used in this work consists of the light curves of objects whose nature was confirmed with spectroscopic or photometric follow-ups, and it is composed by:

(i) cataclysmic variables – CV (461 objects);
(ii) supernovae – SN (536 objects);
(iii) blazar – Bl (124 objects);
(iv) active galactic nuclei – AGN (140 objects);
(v) flare stars – Fl (66 objects);
(vi) RR Lyrae – RRL (292 objects).

2.1 Photometric features

The ability to recognize and quantify the differences between light curves with ML methods, requires many instances of light curves for each class of interest. As extensively discussed (cf. Bloom & Richards 2011; Graham et al. 2012b; Donalek et al. 2013; Wright et al. 2015), in analysing astronomical time series, it is crucial to extract from the light curves a proper set of features. Since light curves are usually unevenly sampled, and not all instances of a certain class are observed with the same number of epochs and S/N ratio, the use of the light curves themselves for classification purposes is therefore challenging, both conceptually and computationally. Therefore, the data need to be homogenized by transforming each light curve into a vector of real-number features generated using statistical and/or model-specific fitting procedures.

In this work, we used the Caltech Time Series Characterization Service (CTSCS), a publicly offered web service (Graham et al. 2012a), to derive from a given light curve a rather complete set of features capable to characterize both periodic (Debosscher et al. 2007; Richards et al. 2011) and non-periodic behaviours.

Among the many possible features provided by the service, we used those listed below.

(i) Amplitude (*ampl*): the arithmetic average between the maximum and minimum magnitude;

$$ampl = \frac{mag_{\max} - mag_{\min}}{2}. \tag{1}$$

(ii) Beyond1std (*b1std*): the fraction of photometric points (≤1) above or under a certain standard deviation from the weighted average (by photometric errors);

$$b1std = P(|mag - \overline{mag}| > \sigma). \tag{2}$$

(iii) Flux percentage ratio (*fpr*): the percentile is the value of a variable under which there is a certain percentage of light-curve data points. The flux percentile $F_{n,m}$ was defined as the difference between the flux values at percentiles n and m. The following flux percentile ratios have been used:

$fpr20 = F_{40,60}/F_{5,95}$
$fpr35 = F_{32.5,67.5}/F_{5,95}$
$fpr50 = F_{25,75}/F_{5,95}$
$fpr65 = F_{17.5,82.5}/F_{5,95}$
$fpr80 = F_{10,90}/F_{5,95}$.

(iv) Lomb–Scargle periodogram (*ls*): the period obtained by the peak frequency of the Lomb–Scargle periodogram (Scargle 1982);

(v) linear trend (*lt*): the slope of the light curve in the linear fit, that is to say the a parameter in the following linear relation:

$$\text{mag} = a * t + b. \tag{3}$$

$$lt = a. \tag{4}$$

(vi) Median absolute deviation (*mad*): the median of the deviation of fluxes from the median flux;

$$mad = median_i(|x_i - median_j(x_j)|). \tag{5}$$

(vii) Median buffer range percentage (*mbrp*): the fraction of data points which are within 10 per cent of the median flux;

$$mbrp = P(|x_i - median_j(x_j)| < 0.1 * median_j(x_j)). \tag{6}$$

(viii) Magnitude ratio (*mr*): an index used to estimate if the object spends most of the time above or below the median of magnitudes;

$$mr = P(mag > median(mag)). \tag{7}$$

(ix) Maximum slope (*ms*): the maximum difference obtained measuring magnitudes at successive epochs;

$$ms = max(|\frac{(mag_{i+1} - mag_i)}{(t_{i+1} - t_i)}|) = \frac{\Delta mag}{\Delta t}. \quad (8)$$

(x) Percent amplitude (*pa*): the maximum percentage difference between maximum or minimum flux and the median;

$$pa = max(|x_{\max} - median(x)|, |x_{\min} - median(x)|). \quad (9)$$

(xi) Percent difference flux percentile (*pdfp*): the difference between the second and the 98th percentile flux, converted in magnitudes. It is calculated by the ratio $F_{5,95}$ on median flux;

$$pdfp = \frac{(mag_{95} - mag_5)}{median(mag)}. \quad (10)$$

(xii) Pair slope trend (*pst*): the percentage of the last 30 couples of consecutive measures of fluxes that show a positive slope;

$$pst = P(x_{i+1} - x_i > 0, i = n - 30, \ldots, n). \quad (11)$$

(xiii) R Cor Bor (*rcb*): the fraction of magnitudes that is below 1.5 mag with respect to the median;

$$rcb = P(mag > (median(mag) + 1.5)). \quad (12)$$

(xiv) Small kurtosis (*sk*): the kurtosis represents the departure of a distribution from normality and it is given by the ratio between the fourth-order momentum and the square of the variance. For small kurtosis, it is intended the reliable kurtosis on a small number of epochs;

$$sk = \frac{\mu_4}{\sigma^2}. \quad (13)$$

(xv) Skew (*skew*): the skewness is an index of the asymmetry of a distribution. It is given by the ratio between the third-order momentum and the variance to the third power;

$$skew = \frac{\mu_3}{\sigma^3}. \quad (14)$$

(xvi) Standard deviation (*std*): the standard deviation of the fluxes.

Table 1. Structure of the confusion matrix for a two classes experiment. The interpretation of the symbols is self-explanatory. For instance, *TP* denotes the number of objects belonging to the class 1 who are correctly classified.

		OUTPUT	
	–	Class 1	Class 2
TARGET	Class 1	*TP*	*FN*
	Class 2	*FP*	*TN*

3 THE METHODS

As it was said before, this work aims to classify transients using a ML approach based on the use of various methods: MLPQNA, RF and KNN.

MLPQNA stands for the classical MultiLayer Perceptron model implemented with a Quasi Newton Approximation as learning rule (Byrd, Nocedal & Schnabel 1994). This model has already been used to deal with astrophysical problems and it is extensively described elsewhere (Brescia et al. 2012; Cavuoti et al. 2014).

RF stands instead for Random Forest, a widely known ensemble method (Breiman 2001), which uses a random subset of data features to build an ensemble of decision trees. Our implementation makes use of the public library scikit-learn (Pedregosa et al. 2011). This method has been chosen mainly because it provides for each input feature a score of importance (rank) measured in terms of its contribution percentage to the classification results.

KNN is the well-known KNN method (Hastie et al. 2001), widely used both for classification and regression. In the case of classification, it tries to classify an object by a majority vote of its neighbours, and the object is then assigned to the most common class among its KNN.

The analysis of the results of the experiments is based on the so-called confusion matrix (Provost, Fawcett & Kohavi 1998), a widely used classification performance visualization matrix, where columns represent the instances in a predicted class, and rows give the expected instances in the known classes. In a confusion matrix defined as in Table 1, the quantities are: *TP*: true positive, *TN*: true negative, *FP*: false positive, *FN*: false negative.

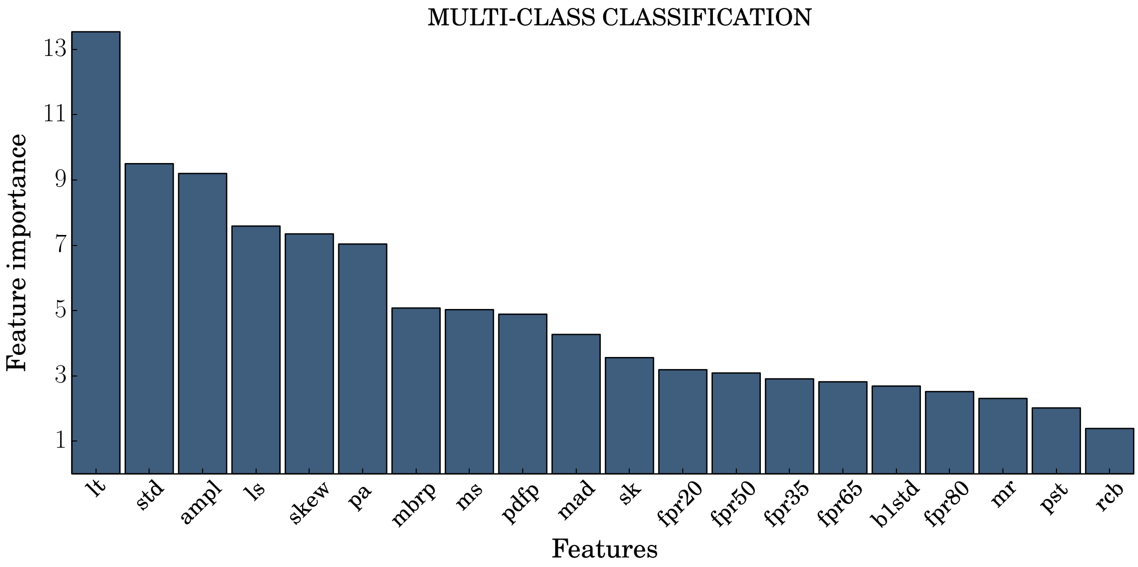
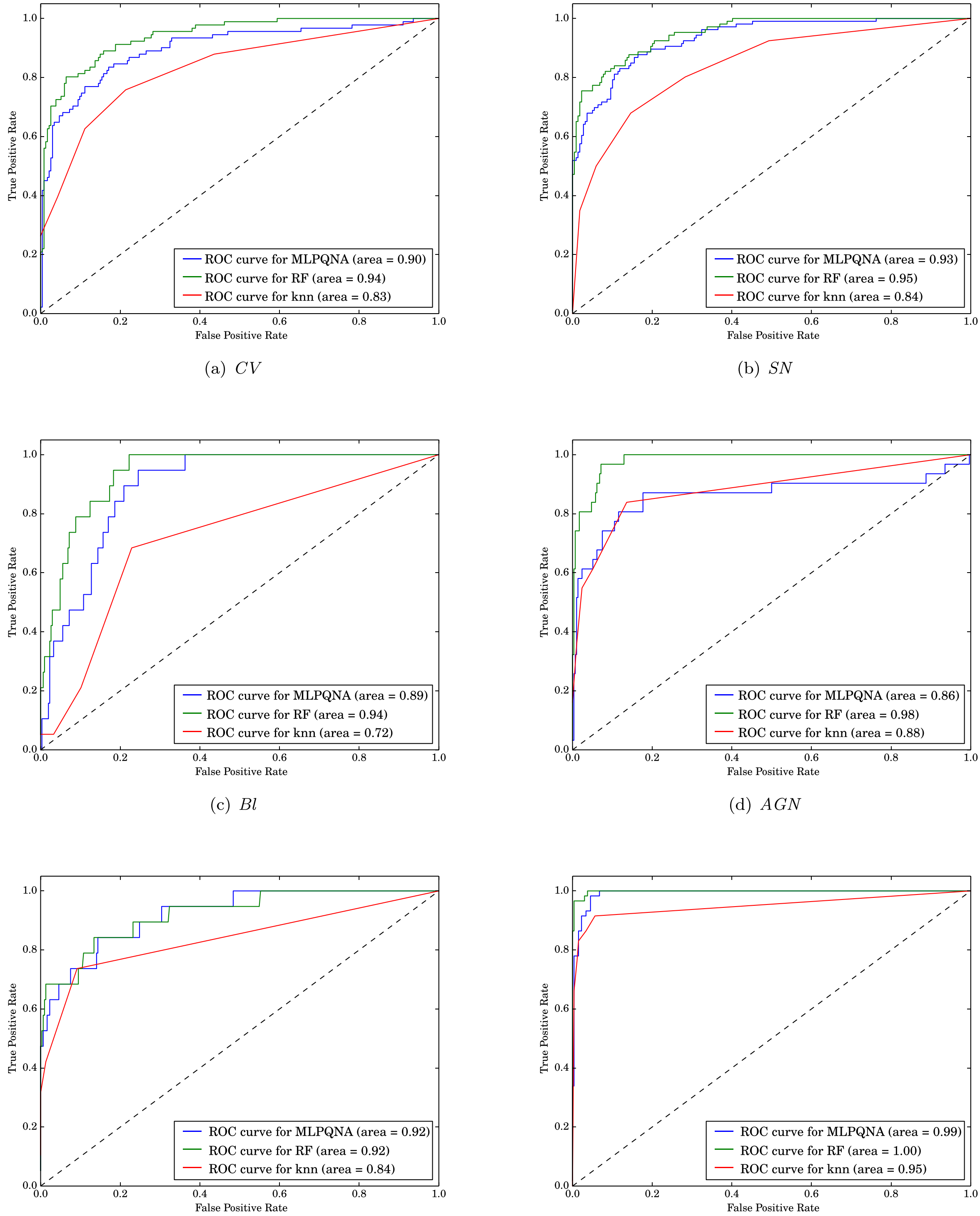


Figure 2. Feature importance list obtained by the RF in the case of the six-class experiment, with the importance percentage for each feature.



(a) *CV*

(b) *SN*

(c) *Bl*

(d) *AGN*

(e) *Fl*

(f) *RRL*

Figure 3. ROC curves for the six-class classification for the three models used. In the case of the KNN model, the curve was obtained by taking into account the limitations imposed by the algorithm, which are determined by the choice of the number of nearest neighbours (in this case, five neighbours induce 20 per cent of quantization).

By combining such terms, it is then possible to derive the following statistical parameters (in brackets the label that will be used in the tables):

(i) overall efficiency (*Eff*): the ratio between the number of correctly classified objects and the total number of objects in the data set;

$$Eff = \frac{TP + TN}{TP + FP + FN + TN}. \tag{15}$$

(ii) class purity (*Pur1* and *Pur2*): the ratio between the number of correctly classified objects of a class and the number of objects

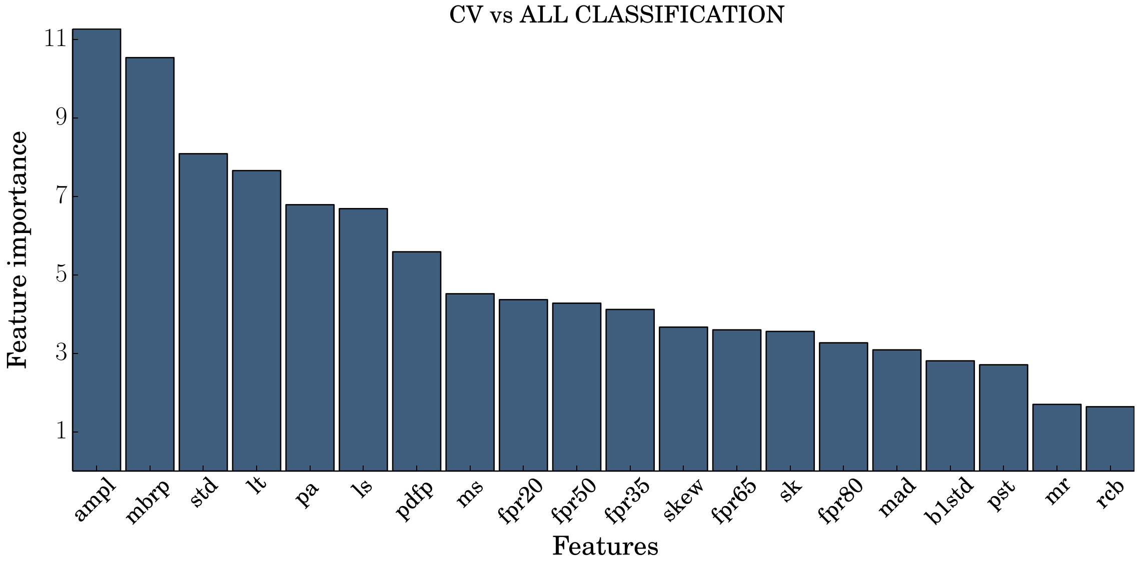


Figure 4. Feature importance list obtained by the RF, with the importance percentage for each feature and for the CV versus ALL classification.

classified in that class, also known as efficiency of a class;

$$Pur1 = \frac{TP}{TP + FP}. \tag{16}$$

$$Pur2 = \frac{TN}{FN + TN}. \tag{17}$$

(iii) class completeness (*Comp1* and *Comp2*): the ratio between the number of correctly classified objects in that class and the total number of objects of that class in the data set;

$$Comp1 = \frac{TP}{TP + FN}. \tag{18}$$

$$Comp2 = \frac{TN}{FP + TN}. \tag{19}$$

(iv) class contamination: it is the dual of the purity. Namely, it is the ratio between the number of misclassified object in a class and the number of objects classified in that class. Since easily derivable from the purity percentages, it is not explicitly listed in the results;

(v) Matthews correlation coefficient (*MCC*): it is an index used as a quality measure for a two-class classification. It takes into account values derived from the confusion matrix, and can be used also if the classes are very unbalanced. It can be regarded as a correlation coefficient between the observed and predicted binary classification, returning a value between −1 and 1. Where −1 indicates total disagreement between prediction and observation, 0 indicates random prediction, and 1 stands for a perfect prediction (Matthews 1975).

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{20}$$

These parameters can be used to describe completely the distribution of the blind test patterns after training.

Moreover, in order to compare the three classifiers used, we also derived the Receiver Operating Characteristic or ROC curve plots for the most significant experiments. An ROC curve is a graphical diagram showing the classification performance trend by plotting the true positive rate against the false positive rate as the classification threshold is varied (Hanley & McNeil 1982). The overall effectiveness of the algorithm is measured by the area under the ROC curve, where an area of 1 represents a perfect classification, while an area of .5 indicates a useless result.

4 CLASSIFICATION EXPERIMENTS

We performed the following classification experiments:

(i) multiclass (six-class), in which the whole catalogue, including all the six classes, was separately considered, in order to investigate the capability to correctly disentangle at once all the given categories of variable objects;

(ii) cataclismic variables (CV) versus ALL, where the category ALL includes AGN, SN, Fl, Bl types. Here, the RRL type was not considered;

(iii) extra-Galactic (AGN and Bl types) versus Galactic (CV, SN and Fl types), to search for an improvement with respect to the previous separation. The inclusion of SN type in the Galactic class is motivated by the fact that, even though mainly observed in external galaxies, they are stars and therefore represent a completely different category with respect to active galactic nuclei;

(iv) SN versus ALL, where ALL includes AGN, Bl, CV, Fl and RRL types.

For each classification experiment, we adopted the same strategy. First of all, we run an RF experiment using all 20 features described in Section 2.1, in order to obtain a feature importance ranking (i.e. the relevance of each feature to the classification expressed in terms of information entropy). The results of the RF experiment allowed us to select different groups of features (ordered by ranking), to be used for a second set of binary classification experiments performed with MLPQNA, RF and KNN. Finally, using the best set of features, we performed an heuristic optimization of the MLPQNA parameters (i.e. complexity of the network topology as well as the Quasi-Newton learning decay factor), aimed at improving the classification results.

We then froze the topology of the MLPQNA using one hidden layer, while for the RF, we chose a 10 000 trees configuration, and finally for the KNN, we chose $k = 5$. Moreover, we always applied a 10-fold cross validation (Geisser 1975), in order to obtain statistically more robust results (i.e. to avoid any potential occurrence of overfitting in the training phase). In terms of performance evaluation, it is important to underline that we were mostly interested to the classification purity percentages. Therefore, these indicators have been primarily evaluated to assign the best results.

4.1 Multiclass

We performed the multiclass classification experiment, to understand the behaviour of the classifiers in the most complex situation, i.e. considering simultaneously all the six available variable object categories. Therefore, as explained above, we performed a preliminary experiment using the RF model with all available input features, thus obtaining the feature importance ranking for this type of classification (Fig. 2). The feature ranking, in fact, is automatically provided by the RF classifier, which assigns a score to all input features, corresponding to their relevance assumed to build the decision rules of the trees during the training phase. Such information indeed is suitable to judge the weight of each individual feature in the decision process and to evaluate its eventual redundancy in terms of contribution to the learning. One useful way to exploit the feature ranking is to engage a training/test campaign, by sequentially adding features to the training parameter space (in order of their importance) and evaluating the training results, until the classification performance reaches a plateau. The final outcome of such campaign is the best compromise between the parameter space dimension and the classification performance. After such preliminary analysis, we then submitted the data set to the RF, MLPQNA and KNN classifiers, by using respectively all, the first five and the first three features of the ranking list in order of importance. A statistical evaluation of the classification results is reported in Tables A4, A5 and A6, while the ROC curves for each class are shown in Fig. 3. From these results it appears evident the worst behaviour of the KNN model with respect to the other classifiers. In terms of class purity, the best behaviour is obtained by the RF model using all available features.

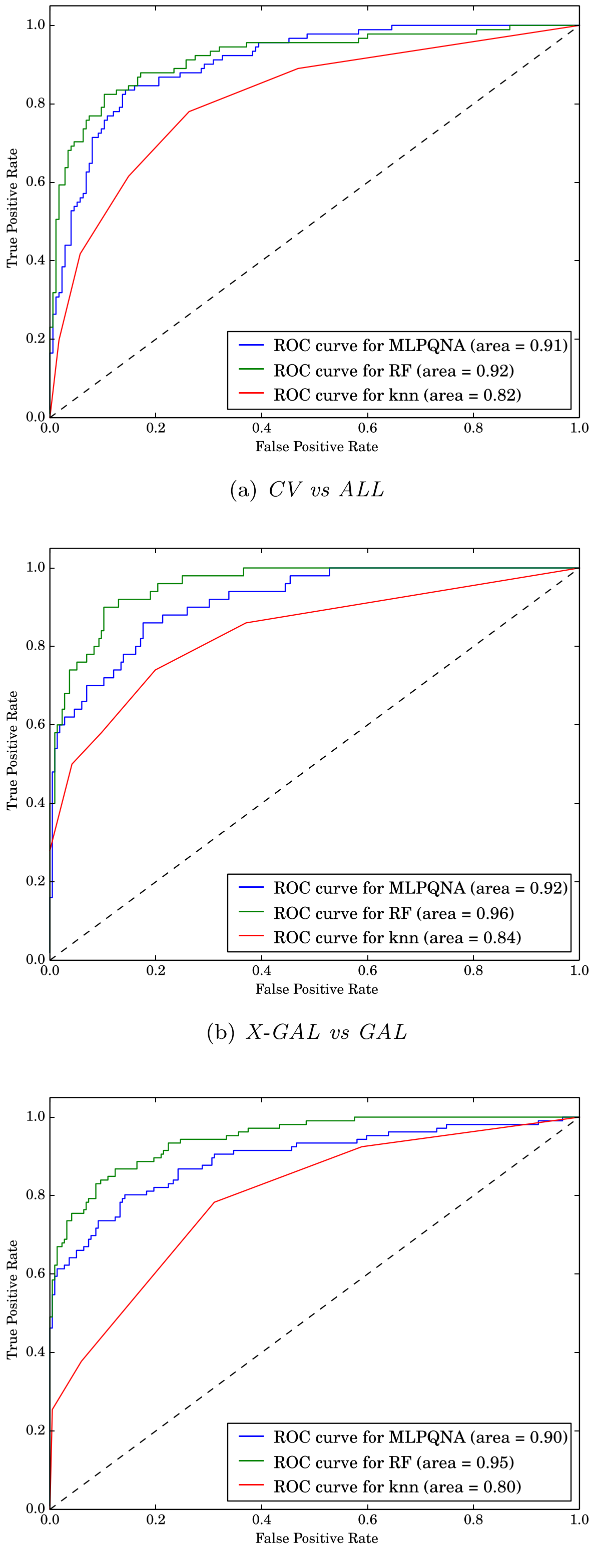
4.2 CV versus ALL

We started by performing an experiment using the RF model and all selected features. The data set was composed by 461 CV and 866 ALL objects. Results are shown in Table A5, while the feature ranking is given in Fig. 4.

Following the feature ranking evaluation strategy, we performed a series of experiments using the MLPQNA, RF and KNN models using different groups of features taken in order of importance: respectively, the first 3, 5, 6, 9, 10, 11 groups and all the 20 features listed in Fig. 4.

In most cases, groups differing by a small number of features (e.g. 5 and 6) led to results with similar performance and, in these cases, we retained as representative the smaller group, assuming that the most of the information is already contained into these groups. Therefore, in the following description of experiments we explicitly report the results only for these relevant cases (see Tables A4, A5 and A6, as well as the related ROC curves in Fig. 5).

From this series of experiments, it appears clear that, regarding MLPQNA, the best configuration is achieved using only five features after the optimization of model parameters (*ampl*, *mbrp*, *std*,



(a) *CV vs ALL*

(b) *X-GAL vs GAL*

(c) *SN vs ALL*

Figure 5. ROC curves for the three different types of classification in the three experiment types. In the case of the KNN model, the curve was obtained by taking into account the limitations imposed by the algorithm, which are determined by the choice of the number of nearest neighbours (in this case five neighbours induce 20 per cent of quantization).

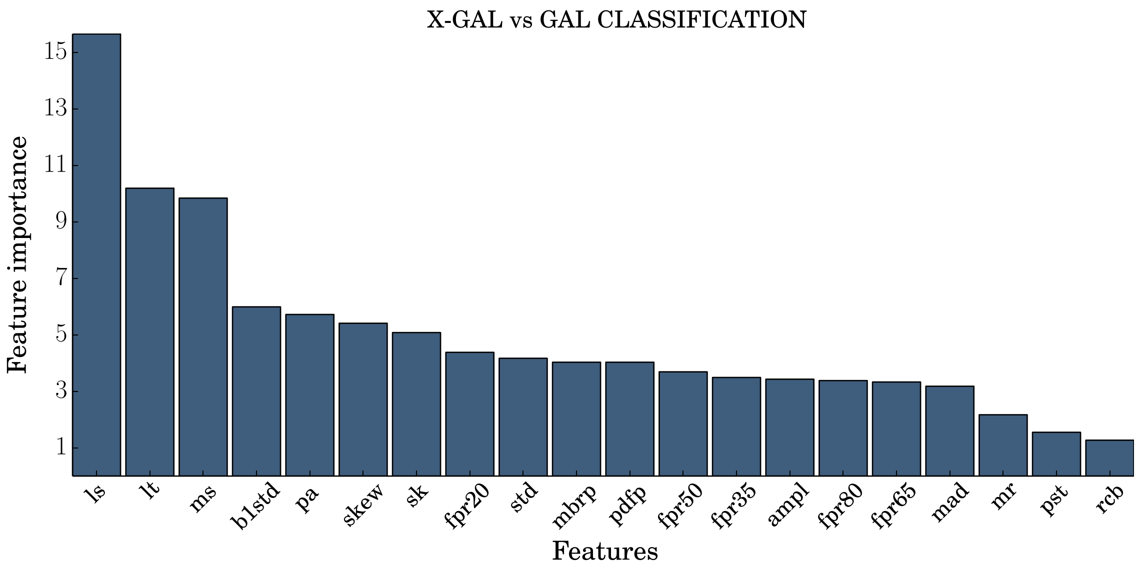


Figure 6. Feature importance list obtained by the RF, with the importance percentage for each feature and for the X-GAL versus GAL classification.

lt and *pa*), while, for the RF, the best results were obtained by retaining all 20 features. Finally, the KNN, which is also the classifier with the worst performance, gives the best result using six features only.

4.3 Extra-Galactic versus Galactic

Also in the case of the classification experiment related to 264 EXTRA-GALACTIC, hereafter called X-GAL, (AGN + Bl as class 1) patterns versus 1063 GALACTIC, hereafter named GAL, (CV + SN + Fl as class 2) patterns, we first performed a feature ranking evaluation with the RF model, by using all available features (see Fig. 6). Again, using the ranking list and the same feature selection strategy described above, we performed a reduced number of experiments using the first 5, 10, and ALL features, by applying all three ML models.

In addition, we performed one additional experiment, using the five features which were selected as most relevant for the CV versus ALL classification case. Results are presented in Tables A7, A8 and A9, while the related ROC curves are shown in Fig. 5. Best classification performance resulted with, respectively, five features for MLPQNA (*ls*, *lt*, *ms*, *b1std* and *pa*) and 10 features for RF and KNN models (*ls*, *lt*, *ms*, *b1std*, *pa*, *skew*, *sk*, *fpr20*, *std*, *mbrp*).

4.4 SN versus ALL

Finally, we performed experiments for SN (class 1), versus ALL (all other classes, labelled as class 2), but in this case we added to the second group also the sixth class containing RR Lyrae, thus obtaining a sample of 536 SN and 1083 ALL class objects. Again, we started from the feature importance evaluation shown in Fig. 7.

As it was already done in the previous cases, we performed the classification experiments with the RF, MLPQNA and KNN models. We report here the results obtained in the cases of, respectively, the first 3, 5 and 10 features in the ranking list. Moreover, we performed additional experiments using the best group of five features obtained from the CV versus ALL experiment (see Fig. 4). Results for the three experiments are reported in Tables A10, A11, A12 and ROC curves in Fig. 5. The best classification performance have been obtained with, respectively, 10 features for RF model (*lt*, *ls*, *pa*, *skew*, *ampl*, *ms*, *std*, *mr*, *fpr20*, *fpr35*) and only 3 features for MLPQNA and KNN classifiers (*lt*, *ls*, *pa*).

5 DISCUSSION

From the experiments previously described, we can notice that, in this context (as imposed by the structure of the parameter space and the size of the data), the RF performs on average slightly better than MLPQNA and objectively better than KNN.

The results presented in the previous paragraph show that at least in presence of such a limited training set the *six-class* experiment is outperformed by the binary classification experiments. The performance achieved by the RF and MLPQNA models for the classes which are more relevant for our work, for instance SNs and CVs categories, led us to investigate two cases of binary classification, respectively, SN versus ALL and CV versus ALL. Furthermore, we approached also the possibility to enclose Blazars and AGN in a single class compared with other categories, thus obtaining a third binary classification experiment, named X-GAL versus GAL. We removed the RR Lyrae category from the binary classification experiments, due to their periodic behaviour, which introduces a very well-defined signature in the data. This has been also derived from the multiclass experiment results, showing how the RR Lyrae objects are easy to classify, thus being not required their inclusion. Only in the case of the SN versus ALL experiment, in order to be as general as possible, we re-introduced the RR Lyrae category.

A first interesting result is that, in spite of the ranking orders obtained for the different experiments and of the results assigned as best, in all cases an accuracy above 80 per cent of efficiency is obtained using the same five most relevant features of the experiment CV versus ALL (*ampl*, *mbrp*, *std*, *lt* and *pa*). This can be understood by comparing the first five positions of the ranking list obtained from the RF for all classification cases, as reported in Figs 4, 6 and 7. In fact, we can notice that among the first five features of Fig. 4, there are two (*lt* and *pa*) in common with other cases, while the two features *ampl* and *ls* are in common between two groups of features (Figs 4 and 7). Moreover, the

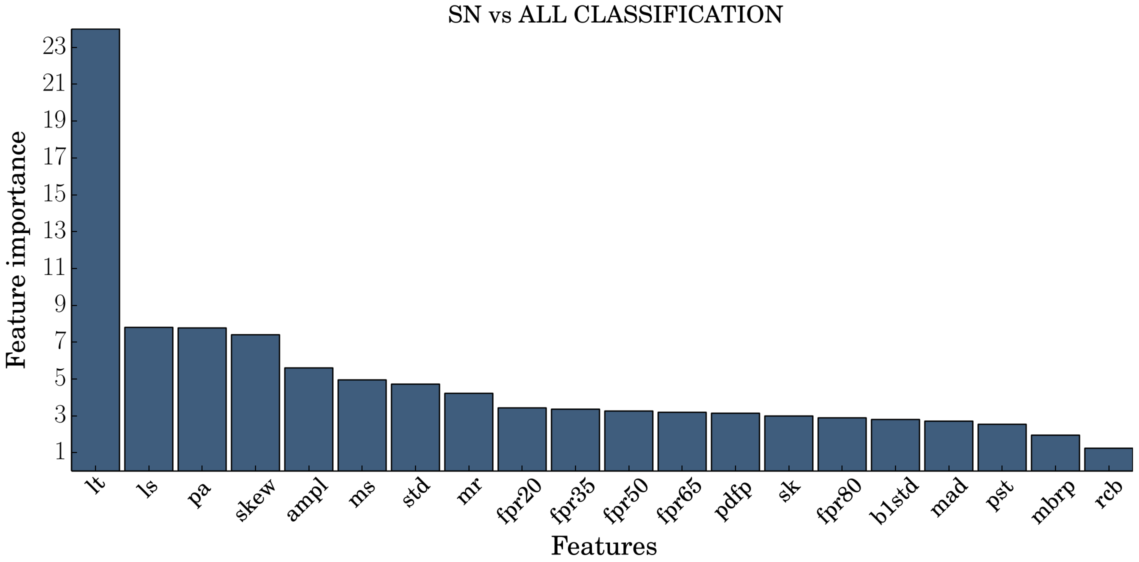


Figure 7. Feature importance list obtained by the RF, with the importance percentage for each feature and for the SN versus ALL classification.

feature *std* is often present within the best groups among different experiments.

Concerning the MCC, this value is almost always above 0.50 for the MLPQNA and RF. In fact, just one experiment shows an MCC below this value, while the best one is 0.74. Therefore, we can conclude that the observed classification with these three classifiers, is close to the expected one, and that the model shows a proper behaviour.

The three classifiers perform differently on different types of objects and, as usual in classification experiments, this implies that the overall performance can be increased by combining the output of the three models. To verify this hypothesis we analysed the overall efficiency variation by taking into account the objects classified by single models and those equally classified by the combination of MLPQNA and RF, MLPQNA and KNN, RF and KNN, and by all three classifiers together.

For this analysis shown in Tables 2, 3 and 4, we performed experiments by randomly splitting the catalogue into a training and a blind test set, containing respectively the 80 per cent and the 20 per cent of the data. The increase in performance is quite evident. These results are also visualized as Venn diagrams in Fig. 8.

The relevance of the various features in the experiments can be better investigated by looking at their distributions. For the sake of clarity in Fig. 9, we show a few relevant examples. In panels a, b and c we show the distribution of the features *lt*, *pa* and *ls* for the SN versus ALL experiment while in panel d and e, we show instead the distribution the parameter *std* in the SN versus ALL and in the CV versus ALL experiments. Finally, in panel f, we show the distribution of the *ampl* feature in the CV versus ALL experiment. In all cases, what appears evident is that individual features fail to separate unequivocally the classes, thus confirming that their combination is needed to achieve a proper classification. Nevertheless, the different roles played by the *std* (panels d and e) in the experiments SN versus ALL and CV versus ALL (cf. Figs 4 and 7, respectively) is confirmed by the histograms.

Table 2. Statistical analysis on the test output for the best experiments of CV versus ALL classification for the three models (5* in Table A4 for the MLPQNA, 20 in Table A5 for the RF, and 6 in Table A6 for the KNN). The first row reports the total amount of test objects. Second, third and fourth rows indicate the overall efficiency obtained by the three models. While the fifth row reports the number of objects equally classified by the three models (i.e. only the objects for which the three models provide the same classification). Finally, the last four rows report the overall efficiencies referred only to the equally classified objects.

CV versus ALL	Size	Fraction
Total test objects	266	–
MLPQNA Eff	224	84 per cent
RF Eff	231	87 per cent
KNN Eff	199	75 per cent
(MLPQNA and RF and KNN) equally classified	189	71 per cent
(MLPQNA and RF) Eff	216	89 per cent
(MLPQNA and KNN) Eff	177	90 per cent
(RF and KNN) Eff	184	90 per cent
(MLPQNA and RF and KNN) Eff	174	92 per cent

Table 3. Statistical analysis on the test output for the best experiments of X-GAL versus GAL classification for the three models (5* in Table A7 for the MLPQNA, 10 in Table A8 for the RF, and 10 in Table A9 for the KNN). The first row reports the total amount of test objects. Second, third and fourth rows indicate the overall efficiency obtained by the three models. While the fifth row reports the number of objects equally classified by the three models (i.e. only the objects for which the three models provide the same classification). Finally, the last four rows report the overall efficiencies referred only to the equally classified objects.

X-GAL versus GAL	Size	Fraction
Total test objects	266	–
MLPQNA Eff	236	89 per cent
RF Eff	243	91 per cent
KNN Eff	224	84 per cent
(MLPQNA and RF and KNN) equally classified	223	84 per cent
(MLPQNA and RF) Eff	233	92 per cent
(MLPQNA and KNN) Eff	211	92 per cent
(RF and KNN) Eff	216	93 per cent
(MLPQNA and RF and KNN) Eff	210	94 per cent

Table 4. Statistical analysis on the test output for the best experiments of SN versus ALL classification for the three models (3* in Table A10 for the MLPQNA, 10 in Table A11 for the RF, and 3 in Table A12 for the KNN). The first row reports the total amount of test objects. Second, third and fourth rows indicate the overall efficiency obtained by the three models. While the fifth row reports the number of objects equally classified by the three models (i.e. only the objects for which the three models provide the same classification). Finally, the last four rows report the overall efficiencies referred only to the equally classified objects.

SN versus ALL	Size	Fraction
Total test objects	325	–
MLPQNA Eff	278	85 per cent
RF Eff	288	89 per cent
KNN Eff	241	74 per cent
(MLPQNA and RF and KNN) equally classified	238	73 per cent
(MLPQNA and RF) Eff	271	90 per cent
(MLPQNA and KNN) Eff	220	89 per cent
(RF and KNN) Eff	229	90 per cent
(MLPQNA and RF and KNN) Eff	218	91 per cent

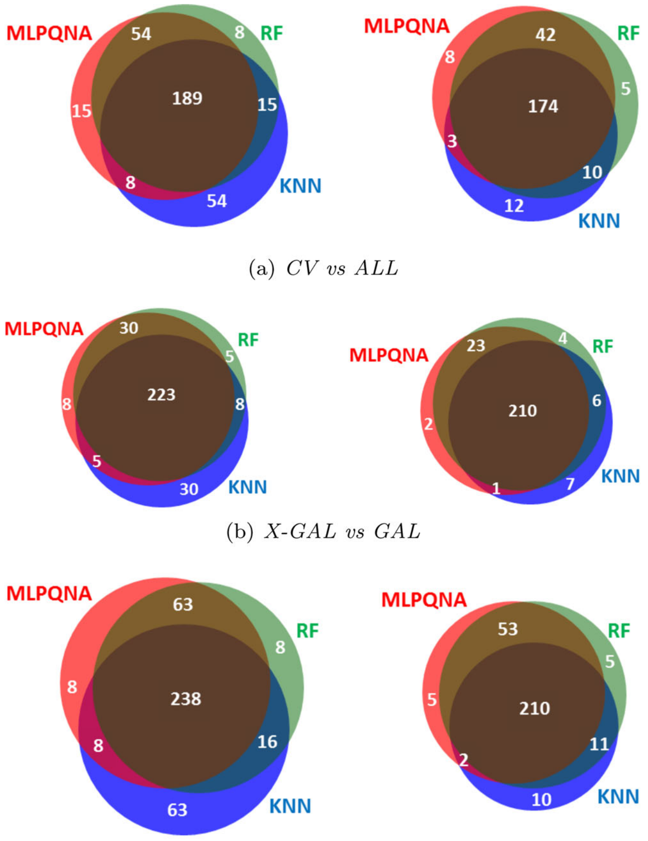
Given the peculiar shape of the SN light curves, it is not a surprise that in the experiment SN versus ALL, the *lt* has a relevance of 24 per cent followed in third position by *pa* with a relevance of 7.7 per cent. The fact that in this experiment the Lomb–Scargle index (*ls*) is ranked second, might seem strange since it is used as an indication of periodic behaviour. The histogram in panel c shows, however, that this is due to the fact that on average objects in the SN class (being non-periodic) have an *ls* much smaller than the ALL class.

In the specific context of the CRTS, a completeness of ∼96 per cent and a purity of 84 per cent in the SN versus ALL classification experiment imply that the sample of candidate SNs produced with our method, would correctly identify ∼2520 out of the 2631 confirmed SNs and would produce a sample of ∼420 possibly spurious objects. These results, however cannot be easily extrapolated to other surveys, since the performance of the method depends drastically on the parameter space covered by the training sample, which as it has been discussed before, is strictly depending on the specific survey.

The capability to disentangle SN class objects through the most relevant selected features appears evident by comparing them among each other. In particular from Figs 10 and 11, it is possible to locate sub-regions entirely populated by *SN*-type objects (those labelled as A in the plots), as well as regions characterized by a weak (labelled as B) or strong (labelled as D) density of SN-type objects. This implies that, besides the particular choice of the classifier, in the parameter space defined by the most relevant features there are combined ranges of feature distributions (for instance, *ls, pa, lt* and *std*) able to classify SN-type objects from the rest of the data types with a high confidence. This evidence is also confirmed by the purity percentages obtained in the case of SN versus ALL experiment by the three classifiers used.

6 CONCLUSIONS

This work focused on the use of three well-tested ML methods, respectively, RF, MLPQNA (Multi Layer Perceptron trained by the Quasi Newton learning rule) and KNN, to classify transient objects and it is a first step towards a framework where different classifiers shall work in collaborative way on the same data to obtain a reliable, accurate and reproducible classification of variable objects.



(a) *CV vs ALL*

(b) *X-GAL vs GAL*

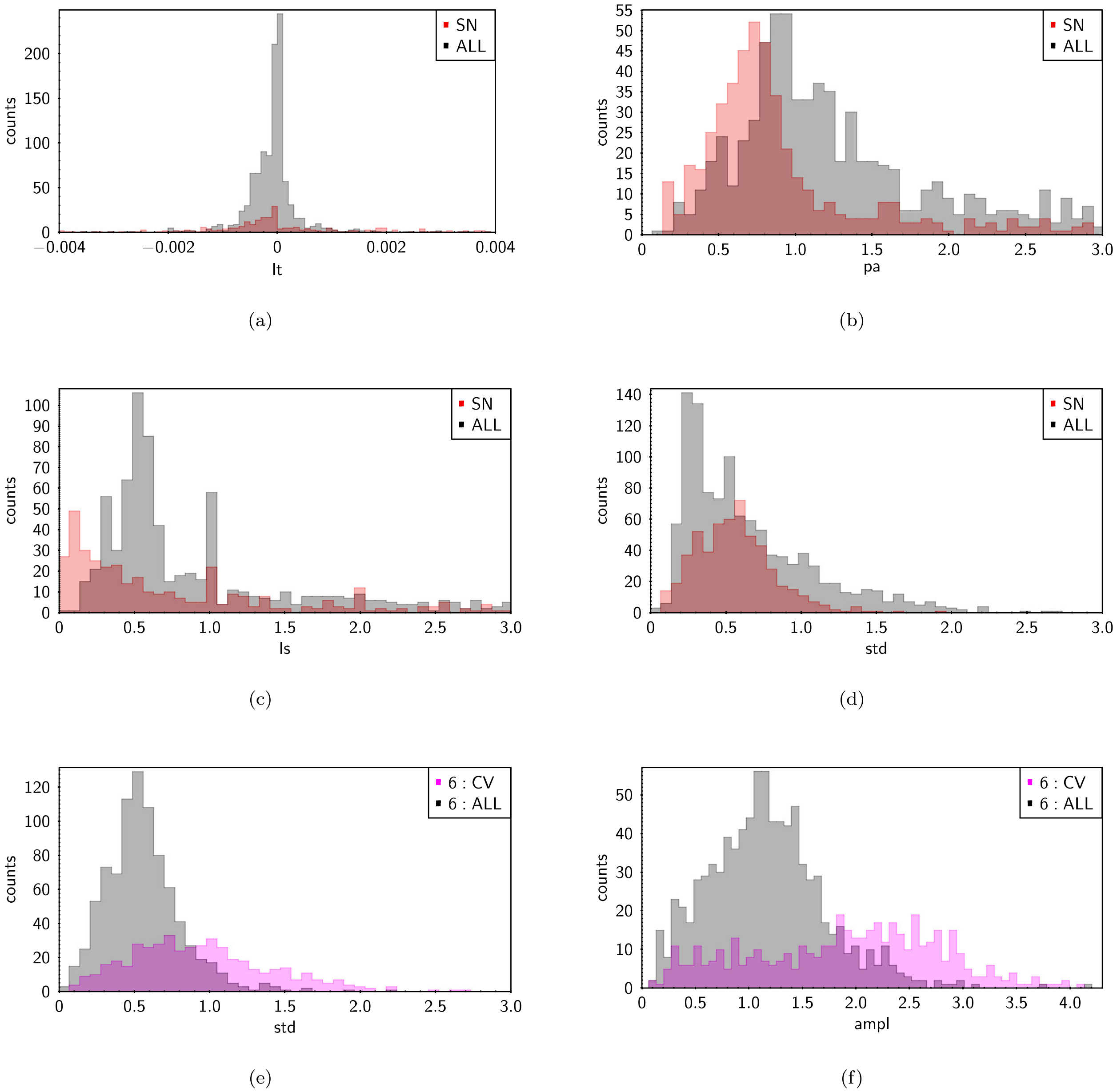
(c) *SN vs ALL*

Figure 8. Venn diagrams showing all the objects (left-hand column) and the correctly classified objects (right-hand column), based on efficiency, for the three different types of classification in the three experiment types. The intersection areas then show the objects that are classified in the same way by different methods. Values are taken from Tables 2, 3 and 4, respectively.

We run a multiclass (all six object categories available) and derived three types of binary classification experiments: (i) CV versus ALL (AGN, SN, Fl, Bl types); (ii) Extra-Galactic (AGN and Bl types) versus Galactic (CV, SN and Fl types); (iii) SN versus ALL (AGN, Bl, CV, Fl and RRL types).

Taking into account the results of the binary classification experiments only, the performance can be summarized as it follows: for the SN versus ALL, the best method is RF, which achieves a ∼87 per cent efficiency, with a completeness of ∼73 per cent and a purity for SNs of ∼86 per cent. In the same experiment, the MLPQNA obtains a slightly higher purity (∼90 per cent) at a price of a lower completeness (∼61 per cent). In the CV versus ALL, the best performance is achieved by the MLPQNA (∼86 per cent efficiency with a completeness of ∼79 per cent and a purity for CVs of ∼80 per cent). It is however worth noticing that the combination of the outcome of the three models allows us to achieve better performance (∼92 per cent efficiency for both experiments). Finally, in the third experiment (X-GAL versus GAL), the best results were achieved by the RF model, obtaining ∼92 per cent efficiency, with a X-GAL class completeness of ∼69 per cent and a purity of ∼88 per cent.

By exploiting the feature importance score provided by the RF model, the ranking between feature grouping and classification performance was investigated and it led to the identification of a special group of features which carry most information, regardless the specific experiment. This is a crucial issue since, in the big data regime



(a)

(b)

(c)

(d)

(e)

(f)

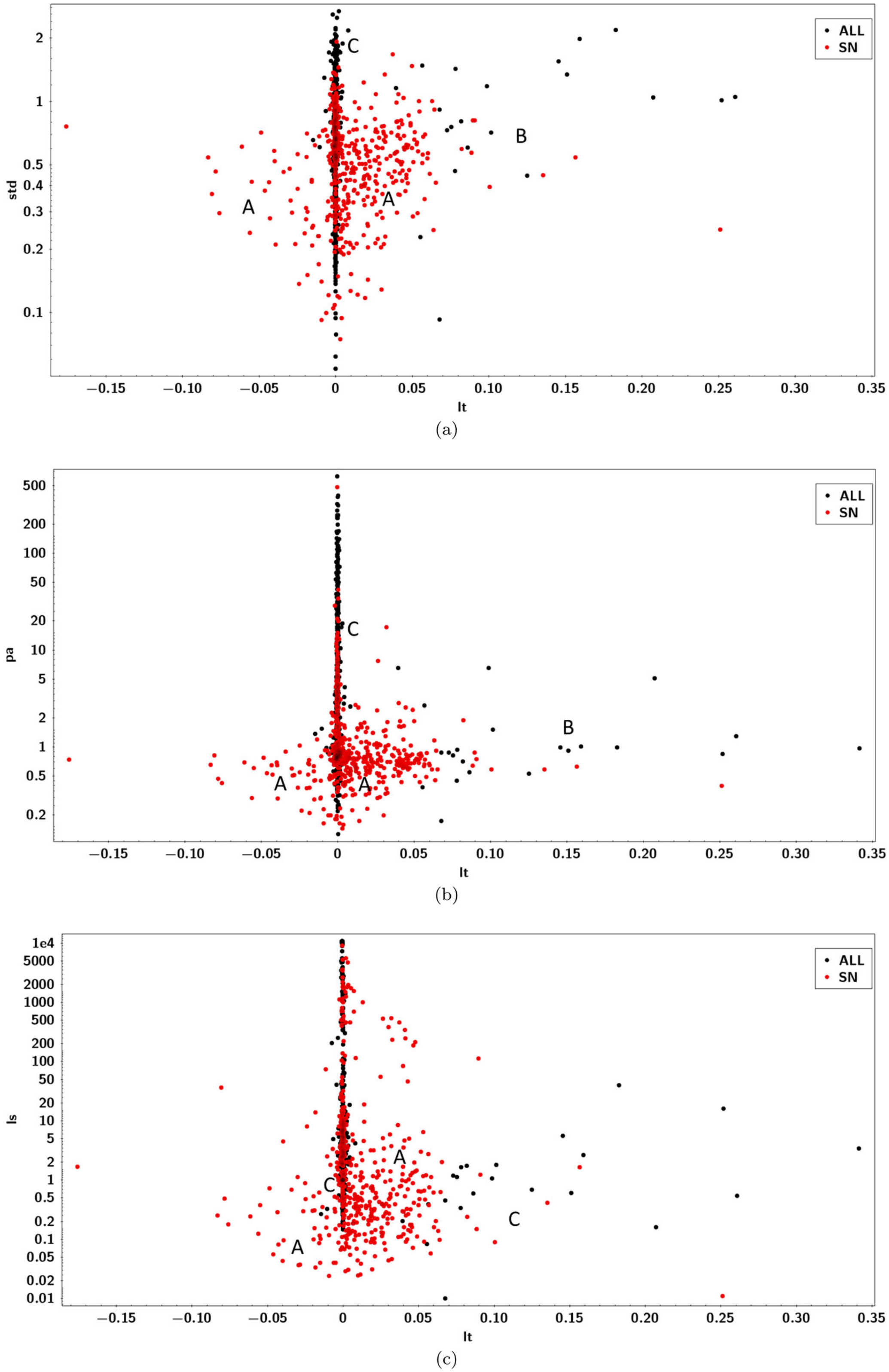
Figure 9. Distribution of the *lt* (panel a), *pa* (panel b), *ls* (panel c) and *std* (panel d) in the case SN versus ALL experiment. The diagram shows a zoomed portion of the distribution to better visualize the region of interest. Red colour is related to SN objects, dark grey colour to ALL class objects, while dark brown shows the overlay area of the histogram. Panels (e) and (f): distribution of the, respectively, *std* and *ampl* features in the case of CV versus ALL experiment. Purple colour is related to CV objects, dark grey represent the ALL class objects, while in dark purple is shown the overlay area of the histogram.

which is typical of future surveys the identification of an optimal set of feature is needed in order to reduce computing time.

Overall, RF and MLPQNA achieve better results when the classifiers are used in combination. The combined and hierarchical use of a wide set of classifiers could be finalized into a framework having as main purpose the capability to disentangle and identify the largest variety of variable objects (Donalek et al. 2013).

ACKNOWLEDGEMENTS

The authors wish to thank the anonymous referee for the very helpful comments and suggestions which contributed to optimize the manuscript. The authors wish to acknowledge financial support from the Italian Ministry of University and Research through the grant PRIN-MIUR 'Cosmology with Euclid' and from the Keck Institute for Space Studies who sponsored the working group on TDA. MB and SC acknowledge financial contribution from the agreement ASI/INAF I/023/12/1. Authors also wish to thank K. Polsterer for useful discussions. This work made use of the CRTS public archive data, of the CTSCS service and of the DAMEWARE infrastructure. CD and SGD acknowledge a partial support from the NSF grants AST-1413600 and AST-1313422. We are thankful to A. J. Drake, A. A. Mahabal, and M. J. Graham for their key contributions in the CRTS project.

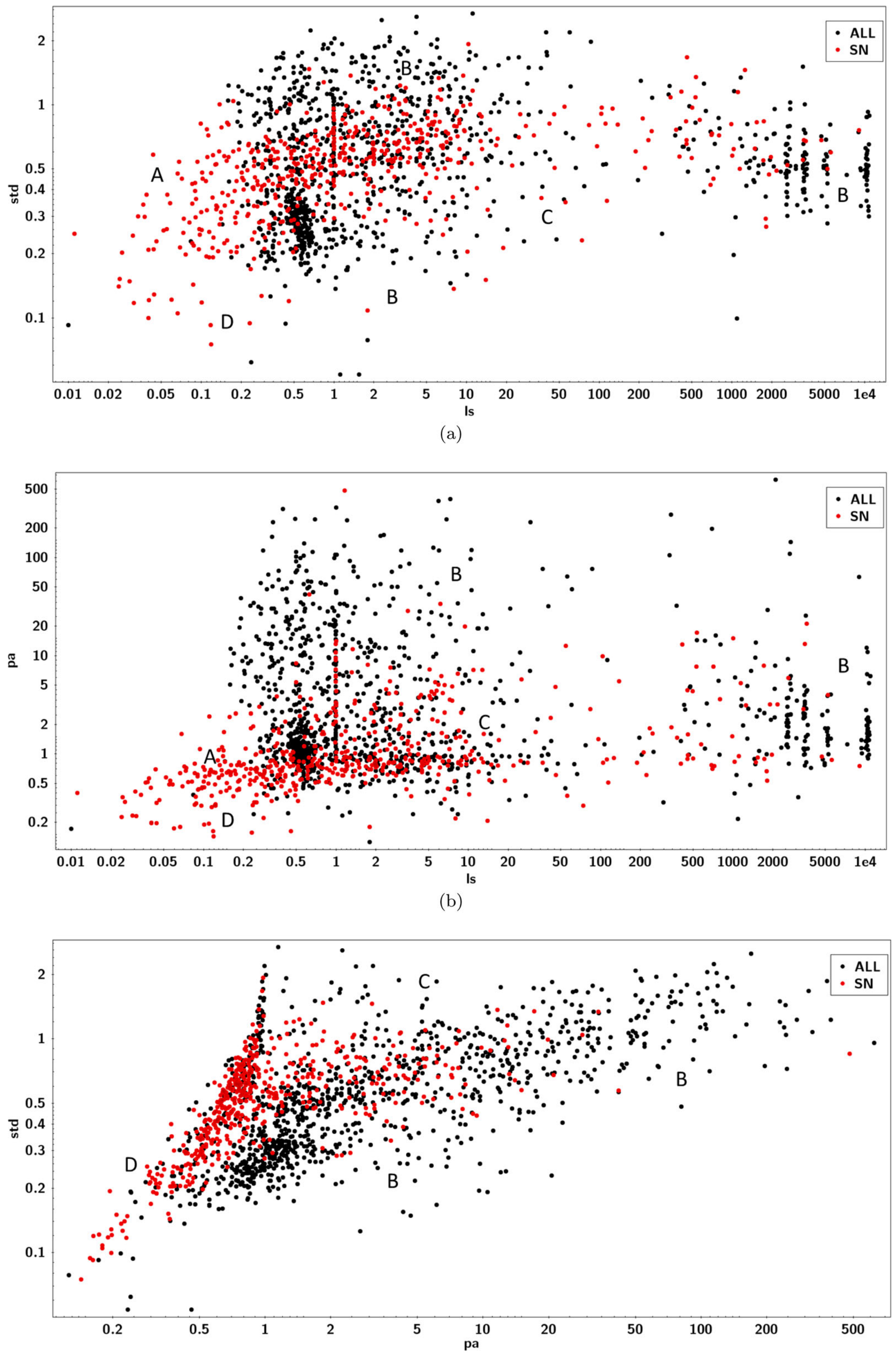


(a)

(b)

(c)

Figure 10. Panel (a): comparison of features *std* versus *lt* in the case of SN versus ALL experiment; panel (b): the same plot but between *pa* and *lt* features; panel (c): the same plot but between *ls* and *lt* features. Red colour is related to SN objects and black to ALL class objects. The labels indicate, respectively, (A) pure SN region (i.e. a region populated only by SN objects), (B) sparse SN region (weak percentage of SN objects), (C) mixed zone and (D) almost pure SN region. The overabundance of points having $lt = 0$ reflects the fact that RRL and AGN as well as any other impulsive variable have in average a constant behaviour.



(a)

(b)

(c)

Figure 11. Panel (a): comparison of features *std* versus *ls* in the case of SN versus ALL experiment; panel (b): the same plot but between *pa* and *ls* features; panel (c): the same plot but between *std* and *pa* features. Red colour is related to SN objects and black to ALL class objects. The labels indicate, respectively, (A) pure SN region (i.e. a region populated only by SN objects), (B) sparse SN region (weak percentage of SN objects), (C) mixed zone and (D) almost pure SN region. The vertical structure at $ls = 1$ is an effect introduced by the sampling frequency of the survey (the structure is mainly populated by AGN, Bl and SN).

REFERENCES

Bloom J. S., Richards J. W., 2011, in Way M. J., Scargle J. D., Ali K. M., Srivastava A. N., eds, Advances in Machine Learning and Data Mining for Astronomy. Chapman & Hall/CRC Press, p. 89

Breiman L., 2001, Mach. Learn., 45, 5

Brescia M., Cavuoti S., Paolillo M., Longo G., Puzia T., 2012, MNRAS, 421, 1155

Brescia M. et al., 2014, PASP, 126, 783

Byrd R. H., Nocedal J., Schnabel R. B., 1994, Math. Program., 63, 129

Castillo E., Gutierrez J. M., Hadi A. S., 1997, Expert Systems and Probabilistic Network Models. Springer-Verlag, New York, p. 605

Cavuoti S., Brescia M., D'Abrusco R., Longo G., Paolillo M., 2014, MNRAS, 437, 968

Cavuoti S., Brescia M., De Stefano V., Longo G., 2015, Exp. Astron., 39, 45

Chang C.-C., Lin C.-J., 2011, ACM Trans. Intell. Syst. Technol., 2, 27:1

Closson Ferguson H., 2015, IAU General Assembly, Meeting 29, 2257590

Debosscher J., Sarro L. M., Aerts C., Cuypers J., Vandenbussche B., Garrido R., Solano E., 2007, A&A, 475, 1159

Donalek C., Graham M., Mahabal A. A., Djorgovski S. G., Drake A. J., Yang M., Maker A., Duan V., 2013, American Astronomical Society, AAS Meeting #221, Automated Classification of Transient and Variable Sources, 352.20

Drake A. J. et al., 2009, ApJ, 696, 870

Drake A. J. et al., 2010, ApJ, 718, 127

du Buisson L. et al., 2015, MNRAS, 454, 2026

Dubath P., 2012, in Sarro L. M., Eyer L., O'Mullane W., De Ridder J., eds, Astrostatistics and Data Mining, Vol. 2, Hipparcos Variable Star Detection and Classification Efficiency. Springer, New York, p. 117

Eyer L., Mowlavi N., 2008, J. Phys. Conf. Ser., 118, 012010

Geisser S., 1975. J. Am. Stat. Assoc., 70, 320

Goldstein D. A. et al., 2015, AJ, 150, 82

Graham M. J. et al., 2012a, Proc. SPIE. Conf. Ser. Vol. 8448, Observatory Operations: Strategies, Processes, and Systems IV. SPIE, Bellingham, p. 8

Graham M. J., Djorgovski S. G., Mahabal A., Donalek C., Drake A., Longo G., 2012b, Data Challegnes of Time Domain Astronomy. Distributed and Parallel Databases 30, Kluwer Academic Publishers, Hingham, MA, p. 371

Graham M. J. et al., 2015, MNRAS, 453, 1562

Hanley J. A., McNeil B. J., 1982, Radiology, 143, 29

Hastie T., Tibshirani R., Friedman J. H., 2001, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, Berlin

Kohonen T., 2001, Self-Organizing Maps, Vol. 30, 3rd edn. Springer, Heidelberg

Lo K. K., Murphy T., Rebbapragada U., Wagstaff K., 2013, Online classification for time-domain astronomy Astroinformaticq11s workshop, IEEE International Conference on Data Mining

McCulloch W. S., Pitts W., 1943, Bull. Math. Biophys., 5, 115

McLachlan G., Peel D., 2000, Finite Mixture Models. John Wiley & Sons, Inc., Hoboken, NJ

Matthews B. W., 1975, Biochim. Biophys. Acta, 405, 442

Pedregosa F. et al., 2011, Mach. Learn., 12, 2825

Provost F., Fawcett T., Kohavi R., 1998, in Fisher D. H., ed., Proc. 15th Int. Conf. Mach. Learn., The Case Against Accuracy Estimation for Comparing Induction Algorithms. Kaufmann Publishers Inc. San Francisco, CA, p. 445

Rebbapragada U., 2014, in Wozniak P. R., Graham M. J., Mahabal A. A., Seaman R., eds, The Third Hot-wiring the Transient Universe Workshop (HTU-III): Data Triage of Astronomical Transients: A Machine Learning Approach. New Mexico, p. 205

Richards J. W. et al., 2011, ApJ, 733, 10

Scargle J. D., 1982, ApJ, 263, 835

Wright D. E. et al., 2015, MNRAS, 449, 451

Yahya S., Bull P., Santos M. G., Silva M., Maartens R., Okouma P., Bassett B., 2015, MNRAS, 450, 2251

APPENDIX A: EXPERIMENT TABLES

Table A1. Results of the experiments with the MLPQNA for the six-class experiment, obtained using the features in order of importance, following the list of Fig. 2. All the results are in percentage.

Statistics	All features	Five features	Three features
Eff	72.46	73.85	73.54
Comp CV	71.43	73.63	72.53
Comp SN	63.21	78.30	80.19
Comp Bl	31.58	26.31	26.82
Comp AGN	61.29	58.06	74.19
Comp Fl	47.37	52.63	57.89
Comp RRL	84.74	96.61	91.52
Pur CV	57.52	71.28	74.16
Pur SN	65.69	76.85	76.58
Pur Bl	33.33	29.41	23.43
Pur AGN	76.00	64.28	58.97
Pur Fl	90.00	83.33	68.75
Pur RRL	87.72	86.36	77.14

Table A2. Results of the experiments with the Random Forest for the six-class experiment, obtained using the features in order of importance, following the list of Fig. 2. All the results are in percentage.

Statistics	All features	Five features	Three features
Eff	79.14	77.30	72.08
Comp CV	79.12	79.12	68.13
Comp SN	83.96	83.96	78.30
Comp Bl	36.84	31.58	26.31
Comp AGN	77.42	67.74	64.52
Comp Fl	52.63	47.37	52.63
Comp RRL	94.91	93.22	93.22
Pur CV	74.22	73.47	66.67
Pur SN	76.72	76.72	74.11
Pur Bl	50.00	37.50	31.25
Pur AGN	85.71	80.77	74.07
Pur Fl	100.00	81.82	71.43
Pur RRL	93.33	94.83	87.30

Table A3. Results of the experiments with the KNN for the six-class experiment, obtained using the features in order of importance, following the list of Fig. 2. All the results are in percentage.

Statistics	All features	Five features	Three features
Eff	55.38	66.77	61.54
Comp CV	64.83	68.13	68.13
Comp SN	62.26	72.64	58.49
Comp Bl	15.79	10.53	5.26
Comp AGN	61.29	61.29	48.39
Comp Fl	10.53	36.84	47.37
Comp RRL	52.54	84.74	76.27
Pur CV	55.14	63.26	59.61
Pur SN	55.46	66.38	65.38
Pur Bl	16.67	10.00	12.50
Pur AGN	70.37	70.37	57.69
Pur Fl	25.00	87.50	52.94
Pur RRL	67.39	89.28	68.18

Table A4. Results of the experiments with the MLPQNA for the CV (class 1) versus ALL (class 2) classification, obtained using the features in order of importance, following the list of Fig 4. All the results are in percentage, except the MCC. The last row (5*) refers to the best result, obtained with an optimization of the model configuration parameters.

Features	Eff	Comp1	Comp2	Pur1	Pur2	MCC
20	77.82	69.23	82.28	67.02	83.72	0.51
3	79.70	54.94	92.57	79.36	79.80	0.53
5	82.71	70.33	89.14	77.11	85.24	0.61
6	79.70	67.03	86.28	71.76	83.42	0.54
9	80.07	73.63	83.43	69.79	85.88	0.56
10	77.82	72.53	80.57	66.00	84.94	0.52
11	79.70	73.63	82.86	69.07	85.80	0.56
5*	86.09	79.12	89.71	80.00	89.20	0.69

Table A5. Results of the experiments with the RF for the CV (class 1) versus ALL (class 2) classification, obtained using the features in order of importance, following the list of Fig. 4 and a cross validation with $k = 10$. All the results are expressed as percentages, except the MCC.

Features	Eff	Comp1	Comp2	Pur1	Pur2	MCC
20	84.02	70.01	91.81	81.96	85.12	0.64
3	77.47	60.49	86.87	71.18	80.46	0.49
5	83.04	71.57	89.38	78.17	85.50	0.62
6	83.49	71.83	89.89	79.21	85.62	0.63
9	84.85	74.46	90.74	81.09	86.84	0.66
10	85.08	74.73	90.86	81.28	86.99	0.67
11	84.32	72.97	90.63	80.55	86.21	0.65

Table A6. Results of the experiments with the KNN for the CV (class 1) versus ALL (class 2) classification, obtained using the features in order of importance, following the list of Fig. 4 and a cross validation with $k = 10$. All the results are expressed as percentages, except the MCC.

Features	Eff	Comp1	Comp2	Pur1	Pur2	MCC
20	78.07	57.75	89.04	73.71	79.82	0.50
3	77.32	59.80	86.94	71.42	80.19	0.49
5	78.07	67.95	83.62	68.91	82.98	0.52
6	79.65	70.04	84.96	71.55	84.00	0.55
9	78.82	64.75	86.51	71.90	82.08	0.52
10	78.75	63.55	87.03	72.33	81.79	0.52
11	78.22	61.87	87.23	72.14	81.10	0.51

Table A7. Results of the experiments with the MLPQNA for the X-GAL (class 1) versus GAL (class 2) classification, obtained using the features in order of importance, following the list of Fig. 6. All the results are in percentage except the MCC. The row (5†) is referred to the features selected in the CV versus ALL experiment. The 5* is the best result obtained by optimizing the model parameters, while the last row (5†*) is the best result obtained in the case of CV versus ALL experiments.

Features	Eff	Comp1	Comp2	Pur1	Pur2	MCC
20	87.97	66.00	93.05	68.75	92.20	0.60
5	88.34	72.00	92.13	67.92	93.43	0.63
10	86.09	72.00	89.35	61.02	93.24	0.58
5†	88.34	66.00	93.52	70.21	92.24	0.61
5*	88.72	68.00	93.52	70.83	92.66	0.62
5†*	88.72	66.00	93.98	71.74	92.27	0.62

Table A8. Results of the experiments with the RF for the X-GAL (class 1) versus GAL (class 2) classification, obtained using the features in order of importance, following the list of Fig. 6, and a cross validation with $k =$ 10. All the results are in percentage except the MCC. The last row (5†) is referred to the features selected in the CV versus ALL experiment.

Features	Eff	Comp1	Comp2	Pur1	Pur2	MCC
20	91.41	66.69	97.64	87.87	92.17	0.71
5	90.73	66.40	96.90	84.50	92.04	0.69
10	91.71	68.51	97.55	88.19	92.58	0.73
5†	88.47	59.91	95.46	76.37	90.64	0.61

Table A9. Results of the experiments with the KNN for the X-GAL (class 1) versus GAL (class 2) classification, obtained using the features in order of importance, following the list of Fig. 6, and a cross validation with $k = 10$. All the results are in percentage except the MCC. The last row (5†) is referred to the features selected in the CV versus ALL experiment.

Features	Eff	Comp1	Comp2	Pur1	Pur2	MCC
20	89.83	71.83	94.44	76.30	92.98	0.68
5	87.80	64.68	93.73	72.19	91.35	0.61
10	90.05	70.33	95.32	78.71	92.59	0.68
5†	86.66	65.09	91.96	67.46	91.39	0.58

Table A10. Results of the experiments with the MLPQNA for the SN (class 1) versus ALL (class 2) classification, obtained using the features in order of importance, following the list of Fig. 7. All the results are in percentage except the MCC. The last column (3*) is referred to the best results obtained by an optimization of the model parameters.

Features	Eff	Comp1	Comp2	Pur1	Pur2	MCC
20	80.00	68.87	85.39	69.52	85.00	0.54
5†	85.23	71.70	91.78	80.85	87.01	0.66
3	84.92	62.26	95.89	88.00	84.00	0.65
5	85.23	72.64	91.32	80.21	87.34	0.66
10	82.15	77.36	84.47	70.69	88.52	0.60
3*	85.23	61.32	96.80	90.28	83.79	0.66

Table A11. Results of the experiments with the RF for the SN (class 1) versus ALL (class 2) classification, obtained using the features in order of importance, following the list of Fig. 7, and a cross validation with $k = 10$. All the results are in percentage except the MCC.

Features	Eff	Comp1	Comp2	Pur1	Pur2	MCC
20	86.60	71.84	93.62	84.47	87.26	0.68
5†	85.98	72.23	92.76	82.67	87.14	0.67
3	85.30	69.74	92.77	82.27	86.26	0.65
5	86.54	72.53	93.27	83.71	87.46	0.68
10	87.34	72.81	94.25	86.00	87.72	0.70

Table A12. Results in percentage, except the MCC, of the experiments with the different groups of features from Fig. 7, obtained using the KNN for the classification SN (class 1) versus ALL (class 2) and a cross validation with $k = 10$.

Features	Eff	Comp1	Comp2	Pur1	Pur2	MCC
20	76.47	57.98	85.92	66.89	80.33	0.45
5†	82.03	63.37	91.21	78.38	83.40	0.58
3	83.32	66.33	91.58	79.38	84.66	0.61
5	79.87	59.39	89.89	74.00	81.81	0.52
10	79.25	65.67	85.81	69.25	83.58	0.52

This paper has been typeset from a TEX/LATEX file prepared by the author.