Alert Classification for the ALeRCE Broker System: The Light Curve Classifier

P. Sánchez-Sáez,[1,2,3] I. Reyes,[1,4,5] C. Valenzuela,[4,1,6,3] F. Förster,[7,1,8] S. Eyheramendy,[3,1] F. Elorrieta,[7,1]
F. E. Bauer,[2,9,1,10] G. Cabrera-Vives,[11,1] P. A. Estévez,[5,1] M. Catelan,[2,9,1] G. Pignata,[12,1] P. Huijse,[13,1]
D. De Cicco,[1,2] P. Arévalo,[14] R. Carrasco-Davis,[5] J. Abril,[15,16] R. Kurtev,[14,1] J. Borissova,[14,1] J. Arredondo,[1]
E. Castillo-Navarrete,[1,4] D. Rodriguez,[1] D. Ruz-Mieres,[1,4] A. Moya,[4,1] L. Sabatini-Gacitúa,[4,1]
C. Sepúlveda-Cobo,[4,1] and E. Camacho-Iñiguez[2]

[1]*Millennium Institute of Astrophysics (MAS), Nuncio Monseñor Sótero Sanz 100, Providencia, Santiago, Chile*
[2]*Instituto de Astrofísica, Facultad de Física, Pontificia Universidad Católica de Chile, Casilla 306, Santiago 22, Chile*
[3]*Faculty of Engineering and Sciences, Universidad Adolfo Ibañez, Diagonal Las Torres 2700, Peñalolén, Santiago, Chile*
[4]*Center for Mathematical Modeling, Universidad de Chile, Beauchef 851, North building, 7th floor, Santiago 8320000, Chile*
[5]*Department of Electrical Engineering, Universidad de Chile, Av. Tupper 2007, Santiago 8320000, Chile*
[6]*Data Observatory, Diagonal Las Torres 2640, Peñalolén, Santiago, Chile*
[7]*Departmento de Matemáticas, Facultad de Ciencia, Universidad de Santiago de Chile, Av. Libertador Bernardo O'Higgins 3663.
Estación Central, Santiago, Chile*
[8]*Departamento de Astronomía, Universidad de Chile, Casilla 36D, Santiago, Chile*
[9]*Centro de Astroingeniería, Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860, 7820436 Macul, Santiago, Chile*
[10]*Space Science Institute, 4750 Walnut Street, Suite 205, Boulder, Colorado 80301*
[11]*Department of Computer Science, University of Concepcón, Edmundo Larenas 219, Concepción, Chile*
[12] *Departamento de Ciencias Fisícas, Universidad Andres Bello, Avda. Republica 252, Santiago, Chile*
[13]*Informatics Institute, Universidad Austral de Chile, General Lagos 2086, Valdivia, Chile*
[14]*Instituto de Física y Astronomía, Facultad de Ciencias, Universidad de Valparaíso, Gran Bretana No. 1111, Playa Ancha, Valparaíso,
Chile*
[15]*European Southern Observatory (ESO), Alonso de Córdova 3107, Vitacura, Santiago, Chile*
[16]*Centro de Estudios de Física del Cosmos de Aragón (CEFCA) - Unidad Asociada al CSIC, Plaza San Juan, 1, E-44001, Teruel, Spain*

ABSTRACT

We present the first version of the ALeRCE (Automatic Learning for the Rapid Classification of Events) broker light curve classifier. ALeRCE is currently processing the Zwicky Transient Facility (ZTF) alert stream, in preparation for the Vera C. Rubin Observatory. The ALeRCE light curve classifier uses variability features computed from the ZTF alert stream, and colors obtained from AllWISE and ZTF photometry. We apply a Balanced Random Forest algorithm with a two-level scheme, where the top level classifies each source as periodic, stochastic, or transient, and the bottom level further resolves each of these hierarchical classes, amongst 15 total classes. This classifier corresponds to the first attempt to classify multiple classes of stochastic variables (including core- and host-dominated active galactic nuclei, blazars, young stellar objects, and cataclysmic variables) in addition to different classes of periodic and transient sources, using real data. We created a labeled set using various public catalogs (such as the Catalina Surveys and *Gaia* DR2 variable stars catalogs, and the Million Quasars catalog), and we classify all objects with ≥ 6 g-band or ≥ 6 r-band detections in ZTF (868,371 sources as of 2020/06/09), providing updated classifications for sources with new alerts every day. For the top level we obtain macro-averaged precision and recall scores of 0.96 and 0.99, respectively, and for the bottom level we obtain macro-averaged precision and recall scores of 0.57 and 0.76, respectively. Updated classifications from the light curve classifier can be found at the ALeRCE Explorer website.

Corresponding author: P. Sánchez-Sáez
pasanchezsaez@gmail.com

1. INTRODUCTION

Brightness variations of astrophysical objects offer key insights into their physical emission mechanisms and related phenomena. In stars, pulsations, both radial and non-radial, can result from a thermodynamic engine operating in their partial ionization layers, when stars are located inside one of the several so-called instability strips that are found in the Hertzsprung-Russell diagram. Eruptive events can be generated by material being lost from a star, or occasionally accreted onto it, as is typical in protostars and young stellar objects (YSOs). Explosive events can occur when material is accreted onto compact objects, such as white dwarfs in the case of cataclysmic variables (CVs) or neutron stars in the case of X-ray binaries, or star mergers. Brightness changes can also originate from the rotation of stars, caused by surface features such as starspots, and/or by stars' ellipsoidal shapes. Finally, eclipses can occur, depending on the observer's line-of-sight, due to the presence of binary companions, planets, and/or other circumstellar material. These and other classes of stellar variability are reviewed and summarized, for instance, in Catelan & Smith (2015), where extensive additional references can be found. In addition, there are a wide array of transients such as kilonovae (Metzger et al. 2010), supernovae (SNe; Woosley et al. 2002), and tidal disruption events, which are beacons of destructive episodes in the life of a star (Komossa 2015). Galaxies, in turn, can also present a wide array of variability phenomena. In those hosting strongly accreting massive black holes (BHs), for instance, variations develop due to the stochastic nature of the accretion disk, corona, and jet emission, potentially related to both the BH properties and the structure of the material in the immediate vicinity (e.g., MacLeod et al. 2010; Caplar et al. 2017; Sánchez-Sáez et al. 2018).

To study the variability of individual objects in detail and use this information to probe different physical models, observations over a wide range of timescales are required. Hence, long and intensive campaigns of a large number of targets are crucial. In recent years surveys covering a significant part of the sky, revisiting the same regions on timescales from days to years, and containing a large sample of serendipitous objects, are now becoming available as predecessors of the Vera C. Rubin Observatory Legacy Survey of Space and Time (LSST; Ivezić et al. 2019).

Among these is the Zwicky Transient Facility (ZTF; Bellm 2014; Bellm et al. 2019), which had first light in 2017 and employs a powerful $47 \, \mathrm{deg}^2$ field-of-view camera mounted on the Samuel Oschin 48-inch Schmidt telescope. ZTF is designed to image the entire northern sky every three nights and scan the plane of the Milky Way twice each night to a limiting magnitude of 20.5 in gri, thus enabling a wide variety of novel multiband time series studies, in preparation for the LSST.

LSST, which aims for first light in 2022, will revolutionize time domain astronomy, enabling for the first time the study of transient and variable objects over long periods of time (~ 10 years) with $\gtrsim 1000$ visits, down to very faint magnitudes ($r \sim 24.5$ for single images of the entire sky every 3 days, ~ 26.1 for yearly stacks, and ~ 27.5 at full depth; 5σ), over a large sky area ($>18,000 \, \mathrm{deg}^2$).

Given the large number of sources that ZTF and LSST will observe (~ 1–40 billion objects), it is critical to develop reliable and efficient variability-based selection techniques. This new information allows us to see through degeneracies which might exist from color characterization alone. These selection techniques should ideally take advantage of the multiband light curves provided by surveys like LSST and ZTF, and separate different subclasses of variable and transient objects without the need for optical spectra, which are still quite expensive to obtain for such large samples.

This new generation of large etendue survey telescopes has demonstrated a growing need for sophisticated astronomical alert processing systems (i.e., systems that are able to detect changes in the sky of an astrophysical origin). These systems involve the real-time processing of data for alert generation, real-time annotation and classification of alerts (up to 40 million events per night) and real-time reaction to interesting alerts using available astronomical resources (e.g., via Target Observation Managers, or TOMs). In order to use these resources intelligently and efficiently, the astronomical community has been developing a new generation of alert filtering systems known as "brokers". One such community broker is the project ALeRCE (Automatic Learning for the Rapid Classification of Events; Förster et al. 2020). ALeRCE is an initiative led by an interdisciplinary and inter-institutional group of scientists from several institutions both in Chile and the United States. The main aim of ALeRCE is to facilitate the study of non-moving variable and transient objects.

ALeRCE is currently processing the ZTF alert stream, providing classifications of different variable and transient objects, in preparation for the LSST era. Two classification models are currently available in the ALeRCE pipeline: a stamp classifier (or early classifier; Carrasco-Davis et al. 2020), that uses a Convolutional Neural Network on the first detection stamp of a source to classify it among five broad classes, namely variable star, active galactic nuclei, SN, asteroid, or bogus; and a light curve classifier (or late classifier), that uses variability features computed from the light curves to classify each source into finer (currently 15) subclasses among three of the five broad classes.

In this work we present the first version of the ALeRCE light curve classifier. This classifier uses several novel features (see Section 3), and employs machine learning (ML) algorithms that can deal with the high class imbalance present in the data, following a two-level scheme. A key goal of ALeRCE is to provide fast classification of transient and variable objects in a highly scalable framework, and thus we only include in this model features that can be computed quickly, avoiding features that require more than one second to compute, based on the computational infrastructure currently at our disposal[1] (for more details see Förster et al. 2020). The main advantage of this classifier is that it can separate multiple classes of transient and variable objects, using features computed from real data, that would be measured from LSST data. Particularly, the light curve classifier can deal with multiple classes of stochastic variable objects (including core, host, and jet-dominated active galactic nuclei, YSOs, and CVs), which have been normally not included by previous classifiers that use real data and classify periodic and transient objects (e.g., Richards et al. 2009; Kim et al. 2014; Nun et al. 2016; Villar et al. 2019a).

This work attempts to separate an unprecedentedly large number of classes (15) of both transients and variable objects using real data (as opposed to using only simulated data). Previous works using real data have mostly focused on selecting either a variety of variable stars classes (e.g., Debosscher et al. 2009; Richards et al. 2012; Kim & Bailer-Jones 2016; Elorrieta et al. 2016; Rimoldini et al. 2019; Hosenie et al. 2019; Zorich et al. 2020), different classes of variable objects, including variable stars and active galactic nuclei (e.g., Kim et al. 2014; Nun et al. 2016), or different classes of transients (Villar et al. 2019a).

To the best of our knowledge three previous works have used real data to classify transients and variable objects, albeit considering a lower number of classes: Martínez-Palomera et al. (2018) used data from the HiTS survey (Förster et al. 2016, 2018) to classify eight transient, active galactic nuclei and variable star classes; Narayan et al. (2018) used data from The Optical Gravitational Lensing Experiment (OGLE; Udalski et al. 1992) and the Open Supernova Catalog (OSC; Guillochon et al. 2017) to classify seven transient and variable star classes; and D'Isanto et al. (2016) used Catalina Real-Time Transient Survey (CRTS; Drake et al. 2009) data to classify six transient and variable object classes. Other works have tested techniques to classify different classes of variables and transients using synthetic data (e.g., Boone 2019), or a combination of synthetic and real data (e.g., Carrasco-Davis et al. 2019).

In addition, this work is the first attempt to separate three different classes of active galactic nuclei (core-dominated or quasi-stellar objects, hereafter "QSO"; host-dominated, hereafter "AGN"; and jet-dominated, hereafter "Blazar"). Previous works have mostly focused on separating active galactic nuclei from the rest (e.g., Butler & Bloom 2011; Peters et al. 2015; Palanque-Delabrouille et al. 2016; Sánchez-Sáez et al. 2019; De Cicco et al. 2019).

The paper is organized as follows. In Section 2 we describe the data used for this work, the procedure for the light curve construction, as well as the taxonomy and the labeled set used to train the classifier. In Section 3 we define the set of features used by the light curve classifier. In Section 4 we describe the different ML algorithms tested for the classifier. In Section 5 we compare the performance of the different models, and report the results obtained for the labeled and unlabeled ZTF sets. Finally in Section 6 we summarize the paper, provide conclusions, and discuss the challenges found during the development of the classifier and the future work.

2. DATA

2.1. *Reference Data*

ALeRCE has been processing the public ZTF alert stream since May 2019, which includes g and r photometry. The ALeRCE pipeline is described in detail by Förster et al. (2020); for clarity, we provide a brief description of the light curve construction process.

The ALeRCE pipeline processes the ZTF Avro alert files.[2] These files contain metadata and contextual in-

---

[1] Using 1 CPU per light curve with r5a.xlarge AWS instances

[2] For details, see https://zwickytransientfacility.github.io/ztf-avro-alert/.
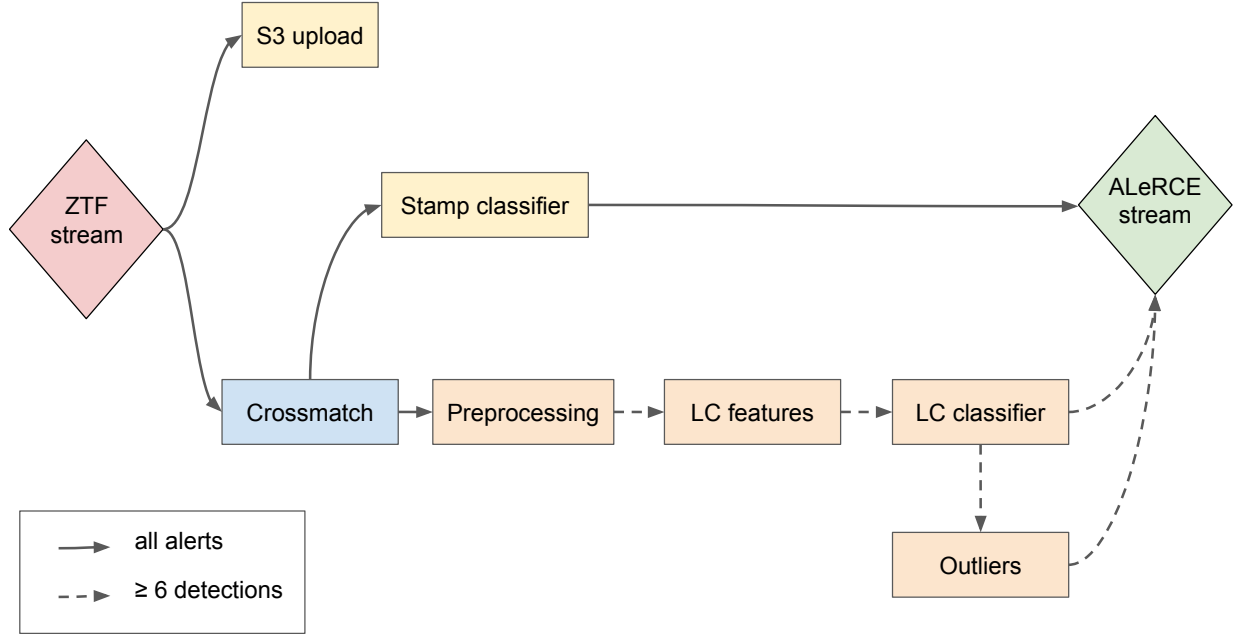
Figure 1. A scheme of the ALeRCE pipeline. ZTF alerts are ingested using Kafka and a series of sequential and parallel steps are initiated. Alerts are stored in AWS S3, classified based on its image stamps, crossmatched with other catalogs, and their photometry corrected to take into account difference fluxes. Aggregated light curves are used to compute basic statistics (for internal use) and, if enough data points exist, features are computed, and a light curve and outlier classifiers are applied before sending an output stream. A PostgreSQL database is populated along the way, which can then be queried.

formation for a single event, which are defined as a flux-transient, a reoccurring flux-variable, or a moving object (Masci et al. 2019). To construct light curves, the ALeRCE pipeline uses: the photometry of the difference-image and reference-image (detections); possible non-detections associated with the target during the previous 30 days of the event (5σ magnitude limit in the difference image based on PSF-fit photometry, called `diffmaglim` by ZTF); the real-bogus quality score reported by ZTF (rb, which ranges from 0 to 1, with values closer to 1 implying more reliable detections); and the morphological classification of the closest object obtained from PanSTARRS1 (Tachibana & Miller 2018). An overview of the pipeline is presented in Figure 1. In summary, the different stages of the pipeline are:

1) Ingestion: the ZTF public stream is ingested using Kafka.

2) S3 upload: the alert Avro packets are stored in AWS S3 for later access.

3) Crossmatch: the position of the alert is used to query external catalogs.

4) Stamp classifier: alerts from new objects are classified using their image cutouts (stamps).

5) Preprocessing: the photometry associated with a given alert is corrected to take into account the use of difference image fluxes (see details below), and simple statistics associated with the aggregated light curve are computed.

6) Light curve features: advanced light curve statistics (features) are computed when there are at least six detections in a given band.

7) Light curve classifier: the light curve classifier described in this work is applied.

8) Outliers: an outlier detection algorithm is applied.

9) ALeRCE stream: the aggregated, annotated and classified light curves are reported in a Kafka stream.

In step 3) we are experimenting with several catalogs, but for this work we use the AllWISE[3] public Source Catalog (Wright et al. 2010; Mainzer et al. 2011), invoking a match radius of 2 arcseconds, to obtain W1, W2, and W3 photometry (using magnitudes measured with profile-fitting photometry, e.g., `w1mpro`).
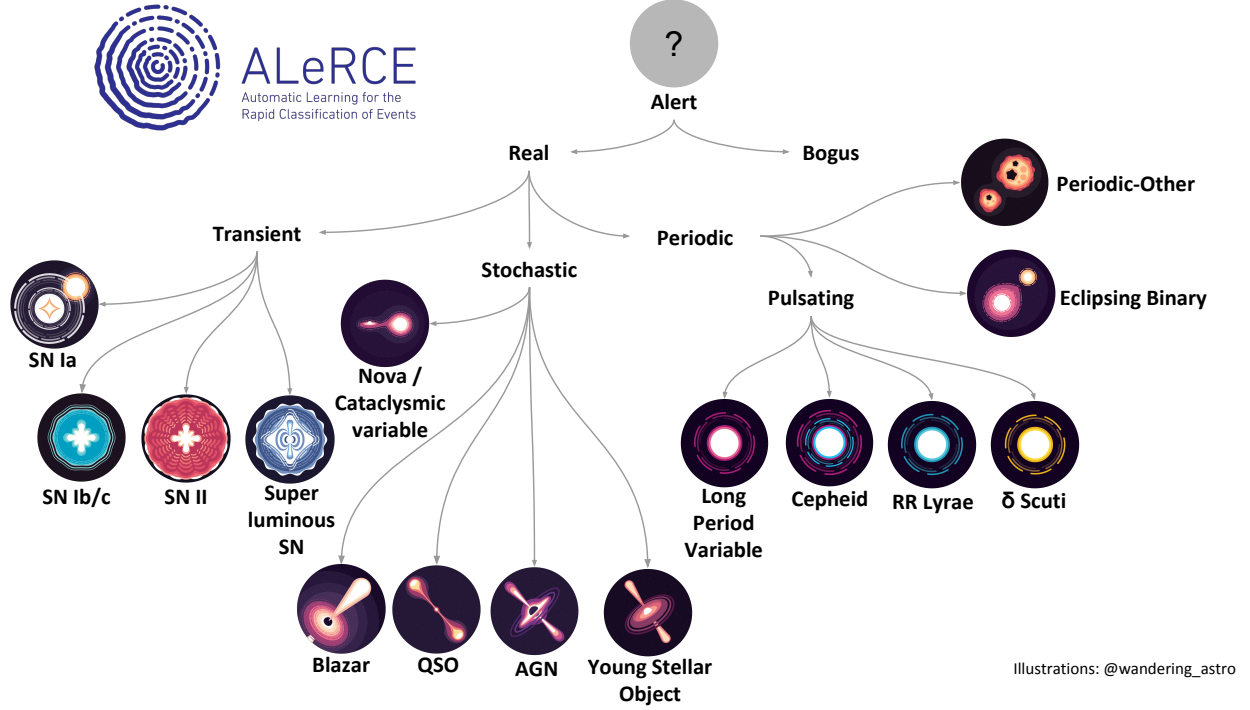
---

[3] http://wise2.ipac.caltech.edu/docs/release/allwise/

Figure 2. Taxonomy tree used in the current version of the ALeRCE light curve classifier.

The preprocessing procedure (step 5) is described in detail in Förster et al. (2020) (see Section A of their appendix). In particular, for the light curve classifier we use the corrected light curves (`lc_corr`; $\hat{m}_{\mathrm{sci}}$ in Förster et al. 2020) for sources whose closest source in the reference image coincides with the location of the alert (in a radius of 1.4 arcseconds). It is important to use the corrected light curves for variable sources, in order to take into account changes in the sign of the difference between the reference and the science images, or possible changes of the reference image. For the rest of the sources, the correction is not possible to perform, and thus we use the light curves obtained using the difference images (`lc_diff`; m_{diff} in Förster et al. 2020), which correspond in general to transient sources. Note that this criteria does not require prior knowledge about the class of the source. Therefore, in this work we use `lc_corr` for sources with available corrected light curves, otherwise we use `lc_diff`, except for the Supernova parametric model features and some optical colors, for which we use `lc_diff` for all the sources (for mode details see Section 3.1 and Appendix A).

2.2. Classification Taxonomy

The first version of the ALeRCE light curve classifier considers 15 subclasses of variable and transient objects, presented as a taxonomy tree defined by the ALeRCE collaboration in Figure 2. The taxonomy is subdivided in a hierarchical fashion according to both the physical properties of each class and the empirical variability properties of the light curves, as follows (in parenthesis we indicate the class name used by the classifier):

- Transient: Type Ia supernova (SNIa), Type Ibc supernova (SNIbc), Type II supernova (SNII), and Super Luminous Supernova (SLSN);

- Stochastic: Type 1 Seyfert galaxy (AGN; i.e., host-dominated active galactic nuclei), Type 1 Quasar (QSO; i.e., core-dominated active galactic nuclei), blazar (Blazar; i.e, beamed jet-dominated active galactic nuclei), Young Stellar Object (YSO), and Cataclysmic Variable/Nova (CV/Nova);

- Periodic: Long-Period Variable (LPV; includes regular, semi-regular, and irregular variable stars), RR Lyrae (RRL), Cepheid (CEP), eclipsing binary (E), δ Scuti (DSCT), and other periodic variable stars (Periodic-Other; this includes classes of variable stars that are not well represented in the labeled set, e.g., sources classified as miscellaneous, rotational or RS Canum Venaticorum-type systems in CRTS).
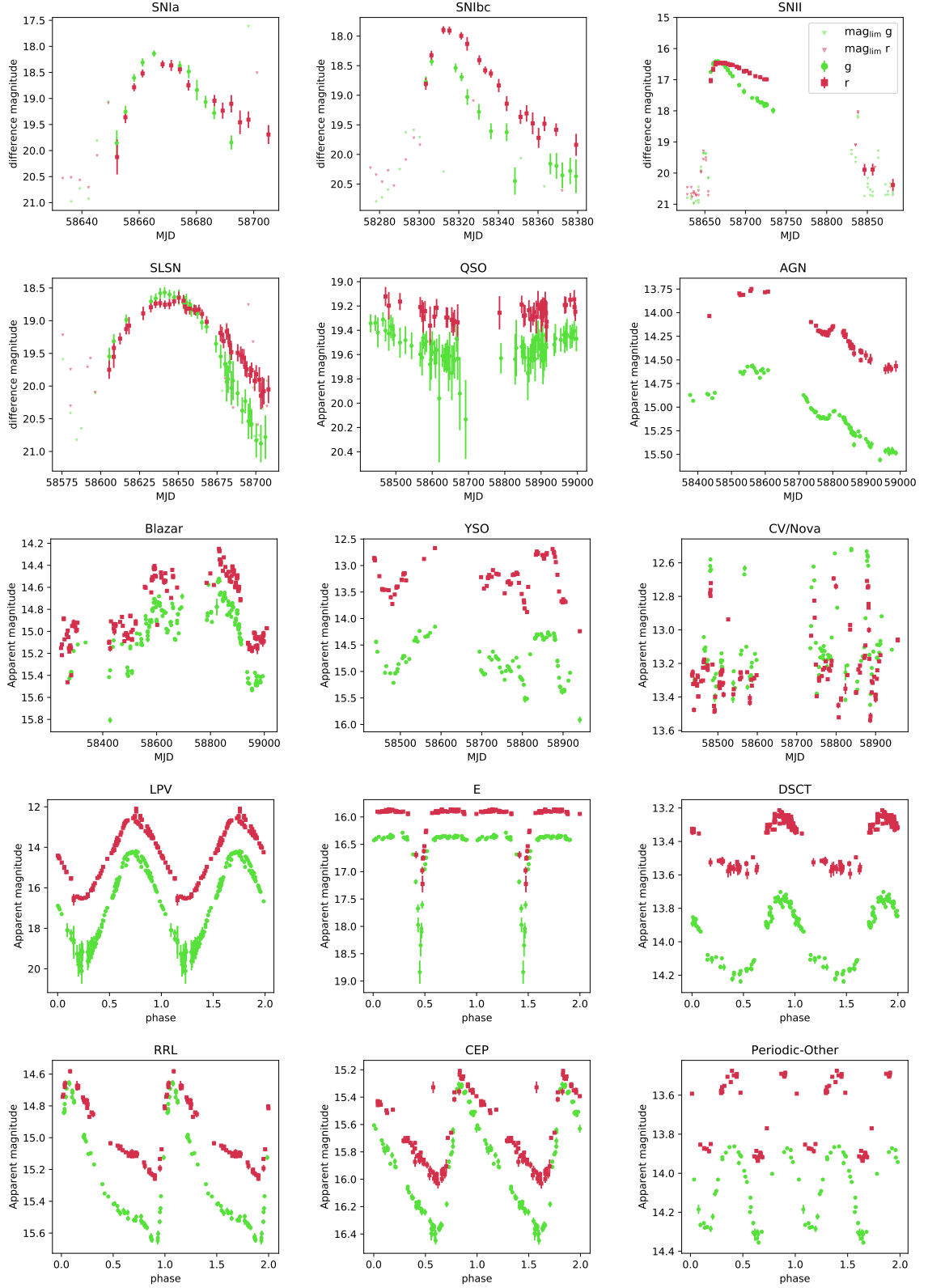
Figure 3. Examples of ZTF light curves of the different classes considered by the light curve classifier. For the transient classes we show the difference magnitude light curves, for the stochastic classes we show the apparent magnitude light curves, and for the periodic classes we show the folded apparent magnitude light curves. Green circles and red squares indicate the g and r bands, respectively. Error bars indicate photometry associated with detections. Triangles denote limiting magnitudes and are shown only for the difference magnitude light curves.

Figure 3 shows examples of light curves of the different classes considered by the light curve classifier, obtained using ZTF data.

It is important to note that there are a number of less common classes which have not been separated out yet in the ALeRCE taxonomy tree, because the number of cross-matched objects in these classes is too low to train a good classification model (e.g. SNe IIb, TDEs, KNe, among others). There is a catch-all "Periodic-Other" class for periodic classes excluded in the taxonomy tree, but not for transient or stochastic classes, and thus, for the moment, these missing classes are being grouped into one or more of the existing ones.
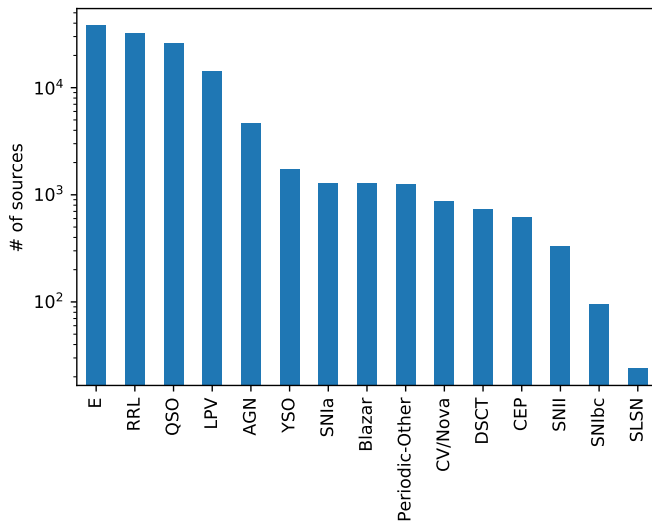
2.2.1. *Labeled Set*

Figure 4. Number of sources per class for the labeled set, as reported in Table 1.

The labeled set (i.e., the set of sources used to define the training and testing sets) for the light curve classifier was built using sources observed by ZTF (i.e., with ZTF light curves), with known labels obtained via spectroscopic and/or photometric analysis by previous works. Further description of the labeled set construction strategy can be found in Förster et al. (2020). We obtained labels from the following catalogs: the ASAS-SN catalogue of variable stars (ASASSN; Jayasinghe et al. 2018, 2019a,b, 2020), the Catalina Surveys Variable Star Catalogs (CRTS; Drake et al. 2014; Drake et al. 2017), LINEAR catalog of periodic light curves (LINEAR; Palaversa et al. 2013), Gaia Data Release 2 (*Gaia*DR2; Mowlavi et al. 2018; Rimoldini et al. 2019), the Transient Name Server database

(TNS)[4], the Roma-BZCAT Multi-Frequency Catalog of Blazars (ROMABZCAT; Massaro et al. 2015), the Million Quasars Catalog (MILLIQUAS, version 6.4c, December 2019; Flesch 2015, 2019), the New Catalog of Type 1 AGNs (Oh2015; Oh et al. 2015), and the SIMBAD database (Wenger et al. 2000). Some additional CV labels were obtained from different catalogs (including Ritter & Kolb 2003), compiled by Abril et al. (2020) (JAbril). It is worth to mention that we only use the labels provided by these catalogs to build our datasets, and not any other information, such as periods, colors or redshifts. A catalog containing the labeled set can be downloaded at Zenodo: 10.5281/zenodo.4279623.

Table 1 lists the number of sources in the labeled set belonging to each class (with their correspondent percentages according to their hierarchical group), and the catalogs from which the classifications were obtained. Only sources with ≥ 6 detections in g or ≥ 6 detections in r were included (considering data obtained until 2020/06/09). It is clear from the table that there is a high imbalance in the labeled set, with some classes representing less than 5% of their respective hierarchical group. Figure 4 shows the (ordered) number of sources per class for the labeled set, and Figure 5 shows the fraction of sources in each class with photometry only in the g band, only in the r band, or in both bands.

3. FEATURES USED BY THE CLASSIFIER

The light curve classifier uses a total of 152 features. We avoid including features that require a long time to compute, for example features that require the use of Markov chain Monte Carlo techniques, since one of the goals of the light curve classifier is to provide a fast and highly scalable classification. 142 of these features are computed using solely the public ZTF g and r data. We excluded the mean magnitude as a feature to avoid that any bias in the labeled set magnitude distribution affects the classification of sources that are fainter (or brighter). Features obtained using the ZTF observed magnitudes are called detection features (56 features in the g band, 56 features in the r band, and 12 multiband features, giving a total of 124 features), and features computed using the ZTF non-detection 5σ magnitude limits `diffmaglim`'s are called non-detection features (nine features for each g and r bands, giving a total of 18 features). These features are described in the following sections (3.1 and 3.2), as well as in Appendix A. Considering the LSST Data Products Definition Docu-

---

Table 1. Labeled set definition

Hierarchical Class	Class	# of sources[†]	Source Catalogs
Transient	SNIa	1272 (74.0%)	TNS
	SNIbc	94 (5.5%)	TNS
	SNII	328 (19.1%)	TNS
	SLSN	24 (1.4%)	TNS
	Total	1718	
Stochastic	QSO	26168 (75.4%)	MILLIQUAS (sources with class "Q")
	AGN	4667 (13.4%)	Oh2015, MILLIQUAS (sources with class "A")
	Blazar	1267 (3.6%)	ROMABZCAT, MILLIQUAS (sources with class "B")
	YSO	1740 (5.0%)	SIMBAD
	CV/Nova	871 (2.5%)	TNS, ASASSN, JAbril
	Total	34713	
Periodic	LPV	14076 (16.2%)	CRTS, ASASSN, $Gaia$DR2
	E	37901 (43.5%)	CRTS, ASASSN, LINEAR
	DSCT	732 (0.8%)	CRTS, ASASSN, LINEAR, $Gaia$DR2
	RRL	32482 (37.3%)	CRTS, ASASSN, LINEAR, $Gaia$DR2
	CEP	618 (0.7%)	CRTS, ASASSN
	Periodic-Other	1256 (1.4%)	CRTS, LINEAR
	Total	87065	

† Values in parentheses correspond to the fraction of sources of a given class (second column) within its corresponding hierarchical class (first column).
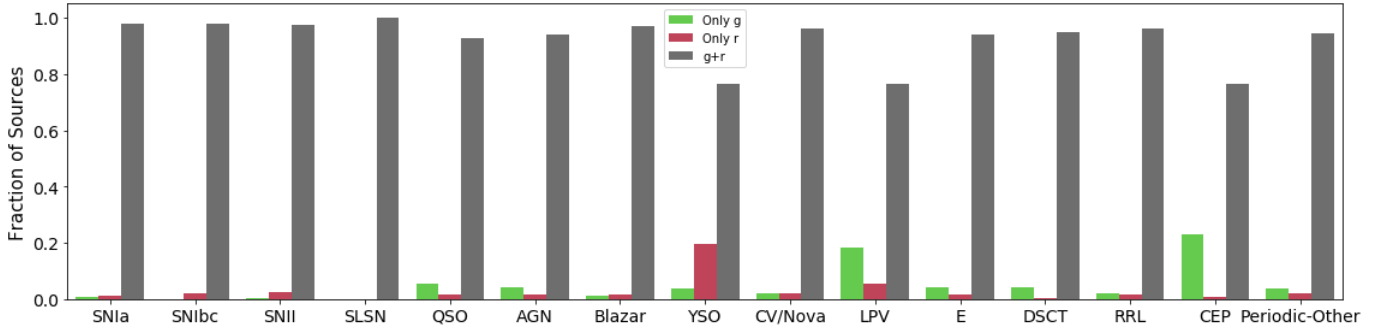

Figure 5. For the sources in the labeled set, this figure shows the fraction of sources in each class with photometry: only in the g band (green); only in the r band (red); or in both bands (grey). The reasons for the non-uniformity of coverage may be physical (strongly red or blue source) or organizational (survey focused on one band only). For most classes, the vast majority of the sources ($\gtrsim$92%) have photometric detections in both g and r; the exceptions are the YSO, LPV, and CEP classes, where only 76% of the sources have photometry in both bands.

ment (Jurić et al. 2019), we expect that all these features would be measured using LSST data.

We also included as features the galactic coordinates of each target (gal_b and gal_l), the W1−W2 and W2−W3 AllWISE colors, and the g−W2, g−W3, r−W2, and r−W3 colors, where g and r are computed as the mean magnitude of the g band and r band light curves for a given source. In addition, we use information included in the Avro files metadata: the sgscore1 parameter, which corresponds to a morphological star/galaxy score of the closest source from PanSTARRS1 (Tachibana & Miller 2018) reported in the ZTF Avro files, with $0 \leq$ sgscore1 ≤ 1, where values closer to 1 imply a higher likelihood of the source being a star; and the median rb (real-bogus) parameter. With these 10 extra features, the total number of features used by the classifier sum to 152.

As we mentioned in Section 2.1, in this work we only consider light curves with ≥ 6 epochs in g or ≥ 6 epochs in r. If a given source has ≥ 6 epochs just in one band, it is included in the analysis, and the features associated with the missing band are considered as -999 values.

This rule applies to all the features used by the classifier; whenever a feature is not available for a given target, we assume a value equal to -999.

3.1. *Detection Features*

Most of the features used by the light curve classifier are computed using the observed magnitudes in the g and r bands (i.e., the detections). There are 56 different features computed for each band, and 12 features computed using a combination of both bands, yielding a total of 124 detection features. The definition of all these features can be found in Table 2. We split the table in three blocks. The first block contains new features defined by this work (i.e., novel features). Some of these features are further described in Section 3.1.1. The second block contains features that correspond to new variants of descriptors included in other works. Some of them are further described in Appendix A. Finally, the third block includes 22 features that come from the Feature Analysis for Time Series (FATS; Nun et al. 2015) Python package. Hereafter, features ending with "_1" are computed for the g band, and features ending with "_2" are computed for the r band, following the notation used in the ZTF Avro files.

Table 2. List of detection features used by the light curve classifier. Features marked with ♦ are computed using both g and r bands at the same time. Features marked with * and ** are further described in Section 3.1.1 and Appendix A, respectively.

Feature	Description	Reference
delta_period	Absolute value of the difference between the Multiband_period and the MHAOV period obtained using a single band	This work
IAR_phi*	Level of autocorrelation using a discrete-time representation of a DRW model	Eyheramendy et al. (2018)
MHPS parameters*	Obtained from a MHPS analysis (three in total)	Arévalo et al. (2012)
positive_fraction	Fraction of detections in the difference-images of a given band which are brighter than the template image	This work
Power_rate* ♦	Ratio between the power of the multiband periodogram obtained for the best period candidate (P) and $2 \times P$, $3 \times P$, $4 \times P$, $P/2$, $P/3$ or $P/4$	This work
PPE* ♦	Multiband Periodogram Pseudo Entropy	This work
$(g\text{-}r)$_max ♦	$g - r$ color obtained using the brightest lc_diff magnitude in each band	This work
$(g\text{-}r)$_max_corr ♦	$g - r$ color obtained using the brightest lc_corr magnitude in each band	This work
$(g\text{-}r)$_mean ♦	$g - r$ color obtained using the mean lc_diff magnitude of each band	This work
$(g\text{-}r)$_mean_corr ♦	$g - r$ color obtained using the mean lc_corr magnitude of each band	This work
delta_mag_fid	Difference between maximum and minimum observed magnitude in a given band	This work
ExcessVar**	Measure of the intrinsic variability amplitude	Allevato et al. (2013)
GP_DRW_tau**	Relaxation time τ from DRW modeling	Graham et al. (2017)
GP_DRW_sigma**	Amplitude of the variability at short timescales ($t << \tau$), from DRW modeling	Graham et al. (2017)
Harmonics parameters**	Obtained by fitting a harmonic series up to the seventh harmonic (14 in total)	(Stellingwerf & Donohoe 1986)
Multiband_period** ♦	Period obtained using the multiband MHAOV periodogram	Mondrik et al. (2015)
Pvar**	Probability that the source is intrinsically variable	McLaughlin et al. (1996)
SF_ML_amplitude**	rms magnitude difference of the SF, computed over a 1 yr timescale	Schmidt et al. (2010)
SF_ML_gamma**	Logarithmic gradient of the mean change in magnitude	Schmidt et al. (2010)
SPM features**	Supernova parametric model features (seven in total)	Villar et al. (2019b)
Amplitude	Half of the difference between the median of the maximum 5% and of the minimum 5% magnitudes	Richards et al. (2011)
AndersonDarling	Test of whether a sample of data comes from a population with a specific distribution	Nun et al. (2015)
Autocor_length	Lag value where the auto-correlation function becomes smaller than Eta_e	Kim et al. (2011)

Beyond1Std	Percentage of points with photometric mag that lie beyond 1σ from the mean	Richards et al. (2011)
Con	Number of three consecutive data points brighter/fainter than 2σ of the light curve	Kim et al. (2011)
Eta_e	Ratio of the mean of the squares of successive mag differences to the variance of the light curve	Kim et al. (2014)
Gskew	Median-based measure of the skew	Nun et al. (2015)
LinearTrend	Slope of a linear fit to the light curve	Richards et al. (2011)
MaxSlope	Maximum absolute magnitude slope between two consecutive observations	Richards et al. (2011)
Meanvariance	Ratio of the standard deviation to the mean magnitude	Nun et al. (2015)
MedianAbsDev	Median discrepancy of the data from the median data	Richards et al. (2011)
MedianBRP	Fraction of photometric points within amplitude/10 of the median mag	Richards et al. (2011)
PairSlopeTrend	Fraction of increasing first differences minus fraction of decreasing first differences over the last 30 time-sorted mag measures	Richards et al. (2011)
PercentAmplitude	Largest percentage difference between either max or min mag and median mag	Richards et al. (2011)
Psi_CS	Range of a cumulative sum applied to the phase-folded light curve	Kim et al. (2011)
Psi_eta	Eta_e index calculated from the folded light curve	Kim et al. (2014)
Q31	Difference between the $3^{\rm rd}$ and the $1^{\rm st}$ quartile of the light curve	Kim et al. (2014)
Rcs	Range of a cumulative sum	Kim et al. (2011)
Skew	Skewness measure	Richards et al. (2011)
SmallKurtosis	Small sample kurtosis of the magnitudes	Richards et al. (2011)
Std	Standard deviation of the light curve	Nun et al. (2015)
StetsonK	Robust kurtosis measure	Kim et al. (2011)

3.1.1. *Description of a relevant set of detection features*

Table 2 summarizes the definitions of the detection features used by the light curve classifier. Some of these features are worth describing in more detail since they are novel features. All other relevant features are described in Appendix A:

- Periodogram Pseudo Entropy: To have an estimate of the confidence of the candidate period (obtained with the multiband MHAOV method), we developed a heuristic based on the entropy of the normalized periodogram peaks, which we denote as periodogram pseudo entropy (PPE). This value is computed by recovering the 100 largest values of the periodogram, normalizing them and computing the entropy of that vector. This feature is computed as

$$PPE = 1 + \frac{1}{\log(100)} \sum_{i=1}^{100} \left(\frac{p_i}{Z}\right) \log\left(\frac{p_i}{Z}\right), \quad (1)$$

  where p_i is the value of the i-th largest peak of the periodogram and $Z = \sum_{i=1}^{100} p_i$. This feature takes values between zero (no clear period stands out) and one (periodogram has a single large peak).

- Power Rate: Ratio between the power of the multiband periodogram obtained for the best period candidate (P) versus the power of the multiband periodogram obtained for n times this period [Power_rate_n $= Power(P)/Power(n \times P)$], where n can take values of 2, 3, 4, 1/2, 1/3, and 1/4. We computed these ratios in order to detect cases where we measure an aliased multiple of the period instead of the true period, which is particularly a common issue for some classes of eclipsing binaries (e.g., Catelan et al. 2013; Graham et al. 2013; McWhirter et al. 2018; VanderPlas 2018, and references therein).

- Irregular autoregressive (IAR) model: Eyheramendy et al. (2018) introduced this model. It is a discrete-time representation of the continuous autoregressive model of order 1 [CAR(1)], which has desirable statistical properties such as strict stationarity and ergodicity without a distributional assumption. The IAR model is defined by

$$y_{t_j} = \phi^{t_j - t_{j-1}} y_{t_{j-1}} + \sigma \sqrt{1 - \phi^{2(t_j - t_{j-1})}} \, \varepsilon_{t_j}, \quad (2)$$

where ε_{t_j} is a white noise sequence with zero mean and unit variance, σ is the standard deviation of y_{t_j}, and $\{t_j\}$ are the observational times for $j = 1, \ldots, n$. We used a modified version of the IAR model, which considers the estimated variance of the measurement errors $\delta_{t_j}^2$ in the likelihood of the model. Thus, by assuming a Gaussian distribution, the negative log-likelihood of the process is given by

$$\ell(\phi, \sigma^2) = \frac{n}{2} \log(2\pi) + \frac{1}{2} \sum_{j=1}^{n} \log \nu_{t_j} + \frac{1}{2} \sum_{j=1}^{n} \frac{e_{t_j}^2}{\nu_{t_j}}, \tag{3}$$

where $e_{t_1} = y_{t_1}$, $\nu_{t_1} = \sigma^2 + \delta_{t_1}^2$, and $\hat{y}_{t_1} = 0$ are the initial values, while $\hat{y}_{t_j} = \phi^{t_j - t_{j-1}} y_{t_{j-1}}$, $e_{t_j} = y_{t_j} - \hat{y}_{t_j}$, and $\nu_{t_j} = \sigma^2 (1 - \phi^{2(t_j - t_{j-1})}) + \delta_{t_j}^2$ for $j = 2, \ldots, n$. Particularly, ϕ describes the autocorrelation function of order 1 for a given light curve. We computed the maximum likelihood estimation of the parameter ϕ (obtained directly from the light curves), and we used this as a feature for our classifier. We denoted this parameter as `IAR_phi`.

- Mexican Hat Power Spectrum (MHPS): Arévalo et al. (2012) proposed a method to compute low-resolution power spectra from data with gaps, where the light curves are convolved with a Mexican hat filter: $F(x) \propto \left[1 - \frac{x^2}{\sigma^2}\right] e^{-x^2/2\sigma^2}$. Gaps, or generally uneven sampling, are corrected for by convolving a unit-valued mask with the same sampling as the light curve and dividing the convolved light curve by it. This method can be used to isolate structures with a characteristic timescale ($t \sim \sigma/\sqrt{2\pi^2}$) in a given light curve, in order to estimate the light curve variance associated with that timescale. We compute the light curve variance at two different timescales of 10 and 100 days. The variance associated with the 10 day timescale ("high" frequency) is denoted `MHPS_high`, while the variance associated with the 100 day timescale ("low" frequency) is denoted `MHPS_low`. We also compute the ratio between the low and high frequency variances for a given band, denoted as `MHPS_ratio`. The logarithm of `MHPS_ratio` is therefore an estimate of the power law slope of the power spectrum of the light curve.

3.2. Non-detection Features

For each detection, the ZTF alert stream includes 5-σ magnitude limits (`diffmaglim`), which are computed from the g and r difference images of the same area of the sky obtained in the previous 30 days, where the target associated with the alert was not detected (non-detections). These non-detections are very informative, since they can, for instance, inform us whether a transient has not been detected before; whether a non-variable source has begun to exhibit a variable behavior; or which range of observed magnitudes we should expect to measure when there are not significant differences between the science and template images, and an alert is not generated. The light curve classifier uses nine different features defined using all the non-detections associated with a given source, computed for both g and r bands, yielding a total of 18 non-detection features. Note that all these features are new, and have not been used before for classification. Table 3 lists the non-detection features used by the light curve classifier. As before, non-detection features ending with "_1" are computed for the g band, and features ending with "_2" are computed for the r band.

4. CLASSIFICATION ALGORITHMS

The labeled set used in this work presents a very high imbalance (see Table 1). For instance, QSOs represent 75.4% of the stochastic sources, while CV/Novae represent just 2.5%. To deal with this issue, we looked for ML algorithms available in the literature that are designed to mitigate the imbalance problem. In particular, we worked with the `imbalanced-learn` Python package (Lemaître et al. 2017). `Imbalanced-learn` includes implementations of several re-sampling algorithms that are commonly used to handle data sets with strong between-class imbalance. The algorithms available in this package are fully compatible with `scikit-learn` methods.

In the following sections we describe the Random Forest (RF) algorithm used by the light curve classifier, as well as other tested ML algorithms. To train each classifier we randomly split the labeled set into a training set (80%) and testing set (20%) in a stratified fashion, preserving the percentage of samples for each class.

4.1. Balanced Random Forest

A Decision Tree (Rokach & Maimon 2008) is a predictive algorithm that uses a tree structure to perform successive partitions on the data according to a certain criterion (e.g., a cut-off value in one of the descriptors or features) and produces possible decision paths, providing a final outcome for each path (the leaves of the tree). Decision Trees are commonly used for classification, where each final leaf is associated with a given class. RFs (Breiman 2001) are algorithms that build multiple Decision Trees, where each tree is trained us-

Table 3. List of non-detection features used by the light curve classifier. Note that "x" stands for either g or r bands.

Feature	Description
`dmag_first_det_fid`	Difference between the last non-detection `diffmaglim` in band "x" before the first detection in any band and the first detected magnitude in band "x"
`dmag_non_det_fid`	Difference between the median non-detection `diffmaglim` in the "x" band before the first detection and in any band the minimum detected magnitude (peak) in the "x" band
`last_diffmaglim_before_fid`	Last non-detection `diffmaglim` in the "x" band before the first detection in any band
`max_diffmaglim_before_fid`	Maximum non-detection `diffmaglim` in the "x" band before the first detection in any band
`max_diffmaglim_after_fid`	Maximum non-detection `diffmaglim` in the "x" band after the first detection in any band
`median_diffmaglim_before_fid`	Median non-detection `diffmaglim` in the "x" band before the first detection in any band
`median_diffmaglim_after_fid`	Median non-detection `diffmaglim` in the "x" band after the first detection in any band
`n_non_det_before_fid`	Number of non-detections in the "x" band before the first detection in any band
`n_non_det_after_fid`	Number of non-detections in the "x" band after the first detection in any band

ing a random sub-sample of elements from a given training set, selected allowing repetition (bootstrap sample of the training set), and using a random selection of features. The final classification is obtained by averaging the classifications provided by each single tree. This average score can be interpreted as the probability (P_{RF}) that the input element belongs to a given class. One of the main advantages of RF is that it naturally provides a ranking of features for the classification, by counting the number of times each feature is selected to split the data.

Chen et al. (2004) proposed a modified RF that can deal with the imbalanced data classification. In their model each individual tree is trained using a sub-sample of the training set that is defined by generating a bootstrap sample from the minority class, and then randomly selecting the same number of cases, with replacement, from the majority classes. `Imbalanced-learn` implements the balanced RF classifier proposed by Chen et al. (2004). For the ALeRCE light curve classifier we use their `BalancedRandomForestClassifier` method, selecting the hyper-parameters (number of trees, maximum number of features per tree, and maximum depth of each tree) with a K-Fold Cross-Validation procedure available in `scikit-learn`, with $k = 5$ folds and using the "macro-recall" as target score (see its definition in Section 5.1).

4.1.1. *The two-level classifier approach*

Considering the hierarchical structure of the taxonomy (see Section 2.2), we decided to construct a balanced RF (BRF) classifier with two-level scheme. The first level consists of a single classifier that separates the sources into three broad classes. The second level consists of three distinct classifiers, which further resolve each class in the first level into subclasses. We then use the probabilities obtained for each independent classifier to obtain the final classification.

In more detail, the first level (top level hereafter) consists of a single classifier which classifies every source as periodic, stochastic, or transient. The second level (bottom level hereafter) consists of three distinct classifiers: Transient, Stochastic, and Periodic. The classes considered by each of these three classifiers are the ones shown in Table 1 and Figure 2. Each classifier in the bottom level is trained using a training subset having only those classes included in the primary top class (for instance, the Transient classifier only includes sources classified as SNIa, SNIbc, SNII, and SLSN). It is important to note that these four classifiers are independent and process the same input features set described in Section 3. The final classification is constructed by multiplying the probabilities obtained for each class of the top level [$P_{top}(transient)$, $P_{top}(stochastic)$, and $P_{top}(periodic)$] with the individual probabilities obtained by their correspondent classifier in the bottom level. Namely, the probabilities of the Transient classifier (P_T) are multiplied by $P_{top}(transient)$, the probabilities of the Stochastic classifier (P_S) are multiplied by $P_{top}(stochastic)$, and the probabilities of the Periodic classifier (P_S) are multiplied by $P_{top}(periodic)$. We denote the product of these probabilities as P. For instance, the probability of a given source being an RRL corresponds to the product of its probability of being periodic (according to the top level) and its probability of being an RRL (according to the Periodic classifier):

$$P(RRL) = P_{top}(periodic) \times P_P(RRL), \qquad (4)$$

while the probability of being a Blazar is computed as:

$$P(Blazar) = P_{top}(stochastic) \times P_S(Blazar). \quad (5)$$

Following this, the sum of the probabilities of the 15 classes for a given source adds up to one. Finally, the class of a given object is determined by selecting the class with the maximum P. Hereafter, we refer to the results presented for the bottom level of the classifier as the final predictions.

The best cross-validation performance was obtained with the following hyper-parameter setting: 500 trees in each classifier, maximum depth trees (the nodes are expanded until all leaves are pure), and a maximum number of features equal to the square root of the total number of features, except for the Stochastic classifier, where we used 20% of the features. In Section 5 we present the results obtained when applying the BRF classifier to the ZTF data.

4.2. *Additional ML algorithms tested*

In addition to RF, we also tested two other supervised classification algorithms: Gradient Boosting and Multilayer Perceptron. These tests were done as a complementary analysis, with the purpose of guiding future efforts in improving the light curve classifier.

None of these methods has a Python implementation particularly designed to handle imbalanced data sets; however, using `imbalanced-learn` we can generate balanced training sets. We present the results obtained using both classifiers in Section 5.

4.2.1. *Gradient Boosting*

Gradient Boosting (GBoost; Friedman 2001) is an ML algorithm that uses an ensemble of weak prediction models (e.g., Decision Trees) to produce a more robust classification. The method implements a boosting algorithm (using a Gradient Descent algorithm) that trains a sequence of weak models, each compensating the weaknesses of its predecessors. `eXtreme Gradient Boosting` (`XGBoost`; Chen & Guestrin 2016) is a package available in several computing languages (including Python) that implements GBoost algorithms for classification and regression in an efficient and scalable way. It has become one of the most used packages for regression and classification in recent years.

For the case of GBoost we followed the same two-level strategy described in Section 4.1.1. However, since the current version of the `XGBoost` multi-class classifier was not designed to deal with highly imbalanced data sets (e.g., Wang et al. 2019), we tested a model that uses `XGBoost` and is trained with a balanced training set. We constructed this balanced training set using the `RandomUnderSampler` and `RandomOverSampler`

methods available in `imbalanced-learn`. For the case of the top level, Periodic, and Stochastic classifiers, we constructed a balanced training set by generating 10 random samples using the `RandomUnderSampler` method, resampling all classes, and concatenating them, in order to obtain a training set with more than 10,000 objects in total for each classifier. For the case of the Transient classifier we used the `RandomOverSampler` method, resampling all classes, to generate one random sample with ~ 600 objects. Each classifier uses the default hyper-parameters defined by the `XGBoost` Python package, with the exception of the boosting rounds, where we used 500, and the objective function, which was set up to do multi-class classification, using the softmax function for multi-class predictions. As in the case of the BRF model, the class of a given object is determined by selecting the class with the maximum probability (obtained by multiplying the probabilities of the top and bottom levels).

4.2.2. *Multilayer Perceptron*

Artificial neural networks (ANNs) are mathematical models inspired by the human brain. ANNs are composed of elemental computational units called neurons (Haykin 1994). ANNs can be used to perform complex tasks such as classification or regression. A Multilayer Perceptron (MLP) corresponds to an ANN whose neurons are ordered by layers, where all neurons belonging to a given layer receive the same input vector and each unit processes this vector independently according to its own parameters. The outputs of all neurons in a layer are grouped and form the input vector for the next layer. For the case of classification, when using the softmax activation function, the final layer provides the probabilities that a given element belongs to a given class. One way of obtaining the final class is to assign the label with the maximum probability in the output layer.

For this model, we also followed the two-level strategy described in Section 4.1.1. We tested different MLP architectures, changing the number of layers and the number of neurons per layer. We used the `Keras API` provided by the Python version of `TensorFlow 2.0` (Abadi et al. 2016). We split the original training set defined above into a new training set (80%) and a validation set (20%). In order to deal with the high imbalance of the training set we used the balanced mini-batches generator for Keras provided by the `imbalanced-learn` Python package. The best performance (considering the categorical cross-entropy loss and accuracy curves for the training and validation sets) was obtained using MLPs with two hidden layers with 256 and 128 neurons for

all the classifiers (top level, Transient, Stochastic, and Periodic). Regularization via the dropout method (Srivastava et al. 2014) is used to prevent overfitting. The dropout fraction is set at 0.50.

5. RESULTS

5.1. *Results for the BRF classifier*

In order to test the performance of our BRF classifier we generated 20 different training and testing sets using the `ShuffleSplit` iterator provided by `scikit-learn`, which uses random permutations to split the data, using each time 80% of the labeled set as training set and 20% as testing set, preserving the percentage of samples for each class in the original labeled set. Then, we trained 20 different BRF models using each training set, and tested their performance using the corresponding testing sets. We emphasize that the testing sets are never used in training their respective models.

Table 4 lists three different scores: precision, recall, and F1-score. For an individual class these scores are defined as:

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i}, \tag{6}$$

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i}, \tag{7}$$

$$\text{F1-score}_i = 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}, \tag{8}$$

where n_{cl} is the total number of classes, TP_i is the number of true positives, FP_i is the number of false positives, and FN_i is the number of false negatives, for a given class i. Despite the high imbalance present in the labeled set, all classes are equally important, and thus we compute macro-averaged scores:

$$\text{Precision}_{\text{macro}} = \frac{1}{n_{cl}} \sum_{i=1}^{n_{cl}} \text{Precision}_i, \tag{9}$$

$$\text{Recall}_{\text{macro}} = \frac{1}{n_{cl}} \sum_{i=1}^{n_{cl}} \text{Recall}_i, \tag{10}$$

$$\text{F1-score}_{\text{macro}} = \frac{1}{n_{cl}} \sum_{i=1}^{n_{cl}} \text{F1-score}_i. \tag{11}$$

For the particular case of the BRF classifier, Table 4 reports the mean and the standard deviation of the macro-averaged scores obtained by the 20 models when applying them to their respective testing sets.

In addition, Figures 6 and 7 show the confusion matrices obtained for the top and bottom levels, respectively. To generate these confusion matrices we used the results

obtained when applying each of the 20 BRF models to their corresponding testing sets, providing for each level the median and 5 and 95 percentiles of the confusion matrices obtained by each of the 20 models.
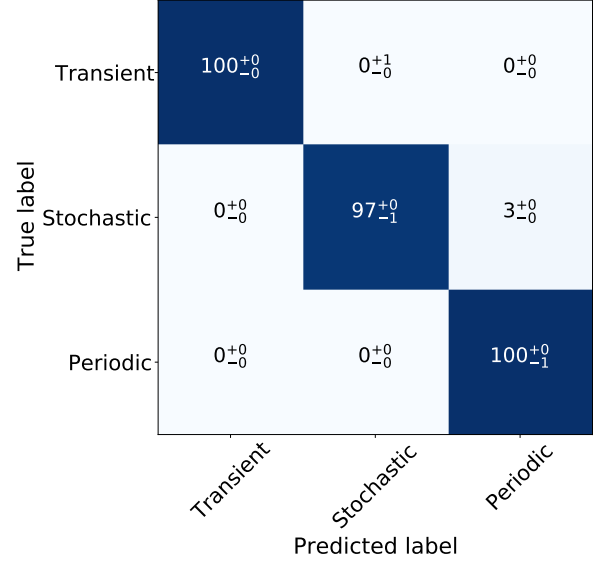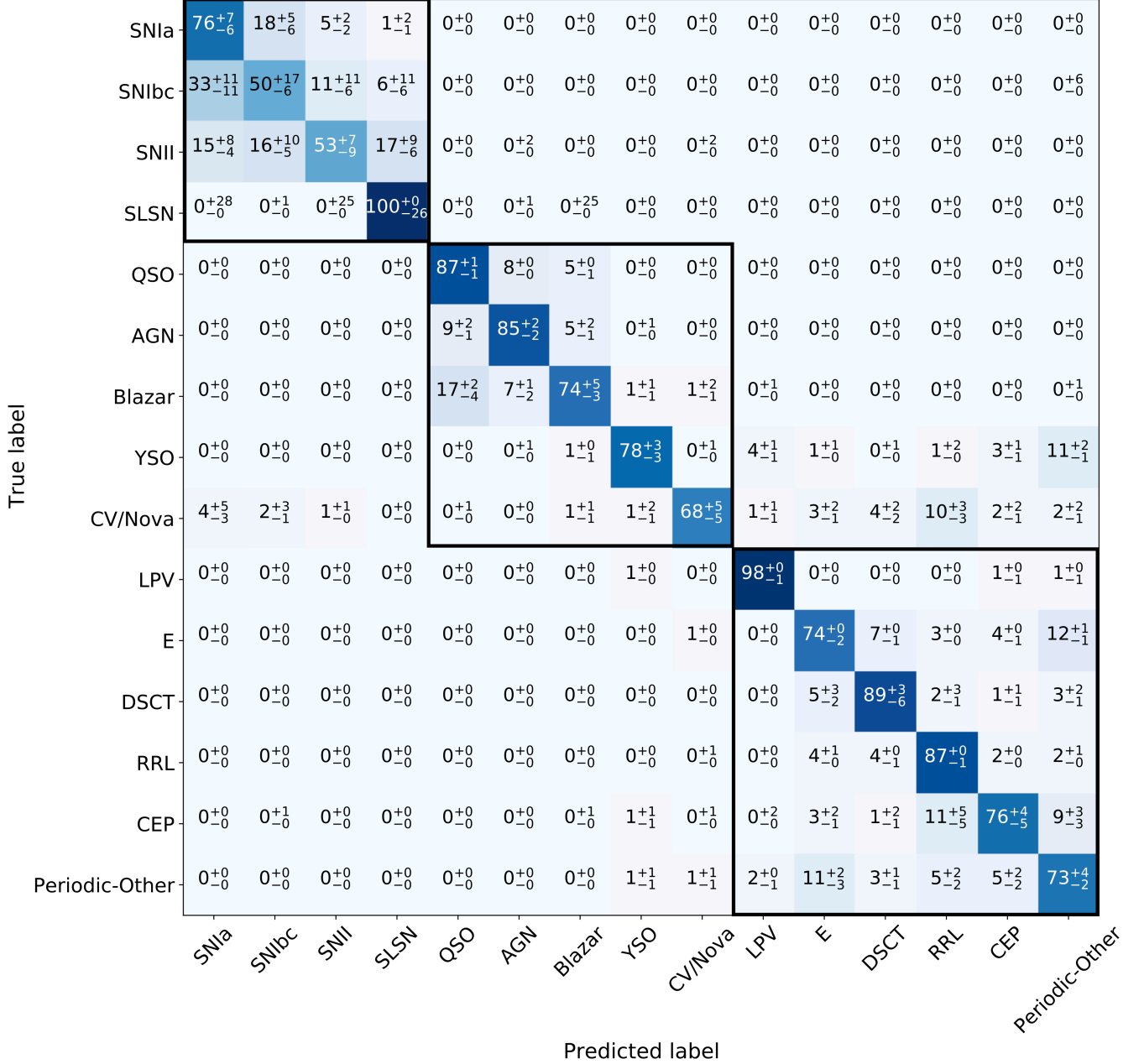

Figure 6. Confusion Matrix for the top level BRF classifier. The confusion matrix was obtained by generating 20 different training and testing sets, and by training 20 independent models using each training set separately. After the training, each model is applied to their respective testing set. We provide the median and 5 and 95 percentiles of the confusion matrices obtained for the 20 testing sets. To normalize the confusion matrix results as percentages, we divide each row by the total number of objects per class with known labels. We round this percentages to integer values. This level shows a high degree of accuracy with a low percentage of misclassifications.

The confusion matrix of the top level shows that the classifier can recover more than 97% of the true labels, and that the contamination between classes is below 3%. The scores obtained reflect the good performance of the top level classifier.

For the case of the bottom level we obtained an F1-score of 0.59, implying significant confusion between classes. From Figure 7 we can see that the fraction of true positives in the confusion matrix of the bottom level has values between 50% and 100%, with mean, median and standard deviation of 78%, 76%, and 14%, respectively. In addition, from the figure it can be observed that the confusion is most often observed among classes with similar characteristics, like among the SN classes (particularly among SNIa versus SNIbc and SNII versus SLSN); among Blazar, AGN, and QSO classes; and among various periodic classes. The highest standard deviation of the predictions is observed for the case of

Table 4. Macro-averaged scores obtained for the BRF classifier in the testing set. The reported values correspond to the mean and standard deviation obtained from the trained 20 models.

Classifier	Precision	Recall	F1-score
BRF - top	0.96 ± 0.01	0.99 ± 0.01	0.97 ± 0.01
BRF - bottom	0.57 ± 0.01	0.76 ± 0.02	0.59 ± 0.01

Figure 7. As in Figure 6, but for the bottom level of the BRF classifier. We provide the median and 5 and 95 percentiles of the confusion matrices obtained for the 20 testing sets. The black squares highlight the three classes of the top level (from top to bottom, transient, stochastic, and periodic, respectively). This matrix is quite diagonal, but shows more misclassification among related subtypes compared to the matrix obtained for the top level.

SLSN. This is a result of the low number of SLSN in the labeled set.

To complement our analysis, in Appendix B we provide the results obtained by a one-level multi-class RF model. From its results we can conclude that the light curve classifier improves considerably when a two-level strategy is followed.

5.1.1. *Comparison with the GBoost and MLP classifiers*

In this work we tested two other classifiers: GBoost and MLP. For these models we present the results obtained by using 80% of the labeled set as a training set, and the remaining 20% as a testing set, preserving the percentage of samples for each class in the original labeled set.

For the case of the GBoost classifier, the macro-averaged precision, recall, and F1-score of the top level have a value of 0.99. For the bottom level the macro-averaged precision, recall, and F1-score are, respectively, 0.72, 0.72, and 0.71. On the other hand, for the MLP classifier the macro-averaged precision, recall, and F1-score of the top level are 0.94, 0.99, and 0.96, respectively. For the bottom level the macro-averaged precision, recall, and F1-score are, respectively, 0.54, 0.69, and 0.58. The confusion matrices obtained for the bottom level of the GBoost and MLP classifiers in the testing set are presented on the left and right sides of Figure 8, respectively.

The precision and F1-score obtained by GBoost are in general better than the ones obtained by the BRF classifier, with the exception of the recall score of the bottom level. However, the fraction of true positives in the confusion matrix of the bottom level range between 5% and 100%, with a mean, median, and standard deviation of 72%, 83%, and 29%, respectively, which explain the lower recall obtained by GBoost, compared to the BRF classifier. In addition, the classes with the largest fraction of true positives in the confusion matrix of GBoost correspond to the most populated classes in the labeled set, like QSOs, which represent 75.4% of the stochastic sources; SNIa, which represent 74.0% of the transients; or LPV, E, and RRL, which represent 16.2%, 43.5%, and 37.3% of the periodic sources, respectively. This is not observed in the results obtained by BRF, where there is no evidence of correlation between the representativity of a given class and its fraction of true positives in the confusion matrix shown in Figure 7.

The results obtained using GBoost are promising. We obtained good scores although the current versions of GBoost available in the literature have not been designed to deal with high imbalance for the case of multi-class classification. Therefore, further efforts should be done to implement GBoost in future versions of the light curve classifier. In particular, new implementations of GBoost for multi-class classification that follow similar approaches to the ones proposed by Chen et al. (2004) or Wang et al. (2019) should be tested, as should combinations of GBoost with data augmentation techniques (i.e., generating synthetic light curves of less populated classes using physical and/or statistical models).

For the case of the MLP classifier the scores obtained are in general lower compared to BRF and GBoost. Its confusion matrix for the bottom level (see Figure 8) presents the same issues already discussed for the case of GBoost. The fraction of true positives in the confusion matrix ranges between 11% and 98%, with a mean, median, and standard deviation of 69%, 74%, and 22%, respectively. Therefore, we conclude that more work should be done in order to obtain better results with MLP.

From these tests we can conclude that the BRF is the model that currently achieves results that are less biased towards the most populated classes in the labeled set, i.e., it is able to predict all sub-classes, and the fraction of true positives does not correlate with how representative a given class in the labeled set is, compared to GBoost and MLP. Thus, we decided to use BRF as the final model for the first version of the ALeRCE light curve classifier. Future work will further exploit the potential of the GBoost and MLP classifiers. For the rest of this paper, presented results correspond to the BRF classifier.

5.1.2. *Results for the BRF classifier excluding AllWISE data*

We tested a version of the BRF classifier that excludes features computed using AllWISE data. The macro-averaged precision, recall, and F1-score of the top level are 0.93, 0.97, and 0.95, respectively. For the bottom level the macro-averaged precision, recall, and F1-score are, respectively, 0.53, 0.72, and 0.55. These scores are slightly smaller than the ones obtained using the original version of the BRF classifier. As can be observed in the confusion matrix shown in Figure 9, the stochastic classes are the most affected by the lack of AllWISE data, particularly YSOs and Blazars, whose fraction of true positives decreased 10% and 16%, respectively. This happens because of the similarities observed between the light curves of these and other classes, which are not easily separated using variability features alone. However, the results obtained by this version of the classifier are still good enough to be used in the case that AllWISE data are not available, as occurs, for instance, for faint objects.
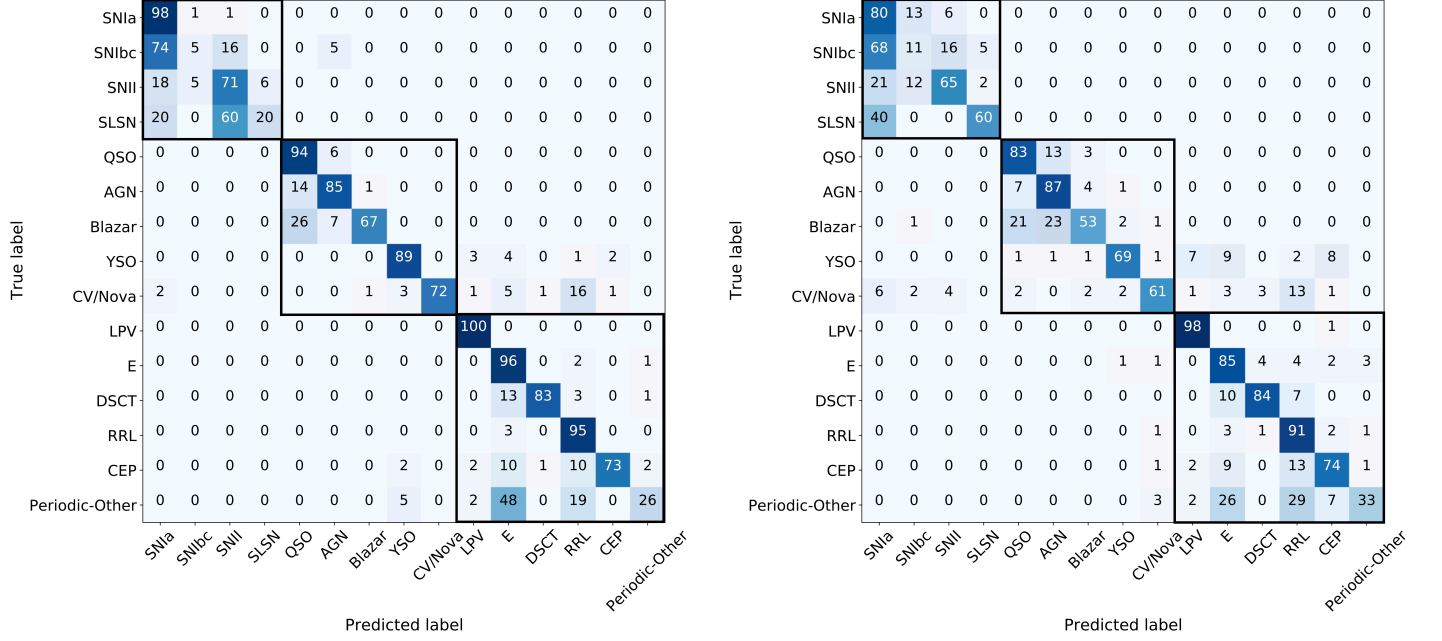
Figure 8. Confusion matrices of the bottom level obtained when applying the GBoost (left) and the MLP (right) classifiers to the testing set. The black squares highlight the three hierarchical classes (from top to bottom, transient, stochastic, and periodic, respectively). To normalize the confusion matrix results as percentages, we divide each row by the total number of objects per class with known labels. We round this percentages to integer values. These matrices present a high percentage of misclassifications compared to the bottom level of the BRF model (Figure 7).

Figure 9. As in Figure 7, but for a model that excludes AllWISE data. We drop the errors, which are comparable to those listed in Figure 7, for simplicity. The results obtained for this model are reasonable, although the classification of some stochastic classes are affected by the lack of AllWISE data.

5.2. *Performance of the BRF classifier as a function of magnitude and number of detections*

As can be seen in Figure 5 for some classes, like YSOs, Cepheids, and LPVs, a non-negligible fraction of sources in the labeled set have photometry available only in one band. It is therefore important to know how well the classifier behaves when a single band is available for a given source.

To evaluate this, we created 20 new testing sets defined considering only those sources with ≥ 6 detections in both g and r bands, from the 20 testing sets previously generated using the `ShuffleSplit` iterator (see Section 5.1). We then classified each new testing set with its respective model, considering: a) the features available for the g and r bands, b) the features available only for the g band (i.e., we hide the features measured using the r band), and c) the features available only for the r band (i.e., we hide the features measured using the g band). Figures 10 and 11 illustrate the results of this analysis. Figure 10 reports the recall values as a function of the average magnitudes per class (i.e., the recall values are computed by comparing the true and predicted labels for the objects within each magnitude bin), with the SN subtypes grouped in the unique class SN. Figure 11 reports the recall values as a function of the number of detections (i.e., the recall values are com-

puted for each bin of the ZTF number of detections). From both figures we can infer that in general the best results are obtained when photometry from both g and r are available, with the exceptions of QSO, CEP and Periodic-Other classes.

From Figure 10 we can also conclude that the reliability of the classification versus the average magnitude is different for each class. These distributions in general follow the magnitude distribution in the labeled set of each class considered in this model (with the exception of RRL). For instance, for the labeled set, the CEP class corresponds to one of the brightest classes, having in general $r < 16$, while the LPV class covers a broader range of magnitudes. On the other hand, from Figure 11, we can infer that in general the classification improves when more detections are available in both bands, with the exceptions of the QSO and Periodic-Other classes.

The results obtained for Periodic-Other are not surprising since this class includes all the periodic classes not considered in the classifier (including several different types of pulsating stars, as well as the rotational variables). The results observed for the CEP class are probably due to the large fraction of Cepheids in the labeled set with photometry only in the g band, which is produced by the saturation limit of the ZTF survey (12.5 to 13.2 magnitudes), and the fact that Cepheids tend to be very bright particularly in the r band, and thus for bright Cepheids the r band light curve is not available. The results obtained for AGNs and QSOs are likely related with incorrect labels, which we discuss further in Appendix C.1.

Peculiar results of the recall values of some classes are reported in Figures 10 and 11 when only one band is available. For SNe, there is a decrease in the recall curve in the g band, presumably due to the fact that in general the g-band light curves of SNe tend to decay faster than the r-band light curves, producing shorter (and thus fewer detections) g-band light curves. This trend can be seen in the SN shown in Figure 3, as well as more generally in the light curve statistics for SNe. The average number of detections of SNe light curves is 12 and 16 in the g and r bands, respectively, and the total time length of SNe light curves corresponds to 53 and 64 days in the g and r bands, respectively. The low recall obtained for bright RRL when only the g band is available may be produced by differences in the variability features measured for different RRL sub-types. This issue is further discussed in Appendix C.2. The zero value recall curves obtained for the CV/Nova class when only the g band is available is produced by the similarities in the AllWISE+ZTF colors of CV/Novae and

some periodic classes. This is discussed in Appendix C.3. Despite this, most of the sources in the labeled set have photometry in both g and r bands (see Figure 5), and thus this low performance obtained for some classes when only one band is available does not highly affect the presented results.

5.3. *The deployed BRF classifier*

In order to use the BRF classifier to classify the ZTF alert stream we need to train a single BRF model. We call this model "the deployed BRF classifier". This "deployed" model corresponds to a totally-independent classifier (i.e., it is different from the previously trained 20 models), and uses the same 152 features presented in Section 3. As in the previous sections, we trained the deployed BRF model using 80% of the labeled set as the training set and the remaining 20% as the testing set. The macro-averaged precision, recall, and F1-score obtained for the top level are 0.96, 0.98, and 0.97, respectively, while for the bottom level they are 0.58, 0.79, and 0.61. The reason for these large differences in the macro-averaged metrics between the top and bottom levels can be understood from Figure 12, which shows the confusion matrix of the bottom level of the deployed BRF model. This confusion matrix is in agreement with the results presented in Figure 7, and shows that some classes, such as the the SNIbc, CV/Nova, YSO, CEP, and Periodic-Other classes, require further work in order to improve the results obtained by the classifier; this may involve data augmentation procedures, and better period estimations, as discussed in the following sections.

5.3.1. *Feature ranking of the deployed BRF classifier*

In Table 5 we list the feature ranking (top 30) for each classifier within the two-level BRF classifier (top level, Transient, Stochastic, and Periodic). The feature ranking is computed considering which features separate better the subclasses within each classifier, with more informative features having higher ranks (for more details see Hastie et al. 2009). From the table we can see that for all the classifiers, a considerable fraction of the top 30 features correspond to colors computed using the AllWISE and ZTF photometry, as well as new detection features (i.e., features not included in the FATS package) and non-detection features. Moreover, it can be observed that the ranking of features changes for each classifier.

The top level classifier is dominated by different types of features: ZTF and ALLWISE colors, morphological properties of the images (`sgscore1`), variability features related with the amplitude of the variability at short and long timescales (`MHPS_low`, `GP_DRW_sigma`,
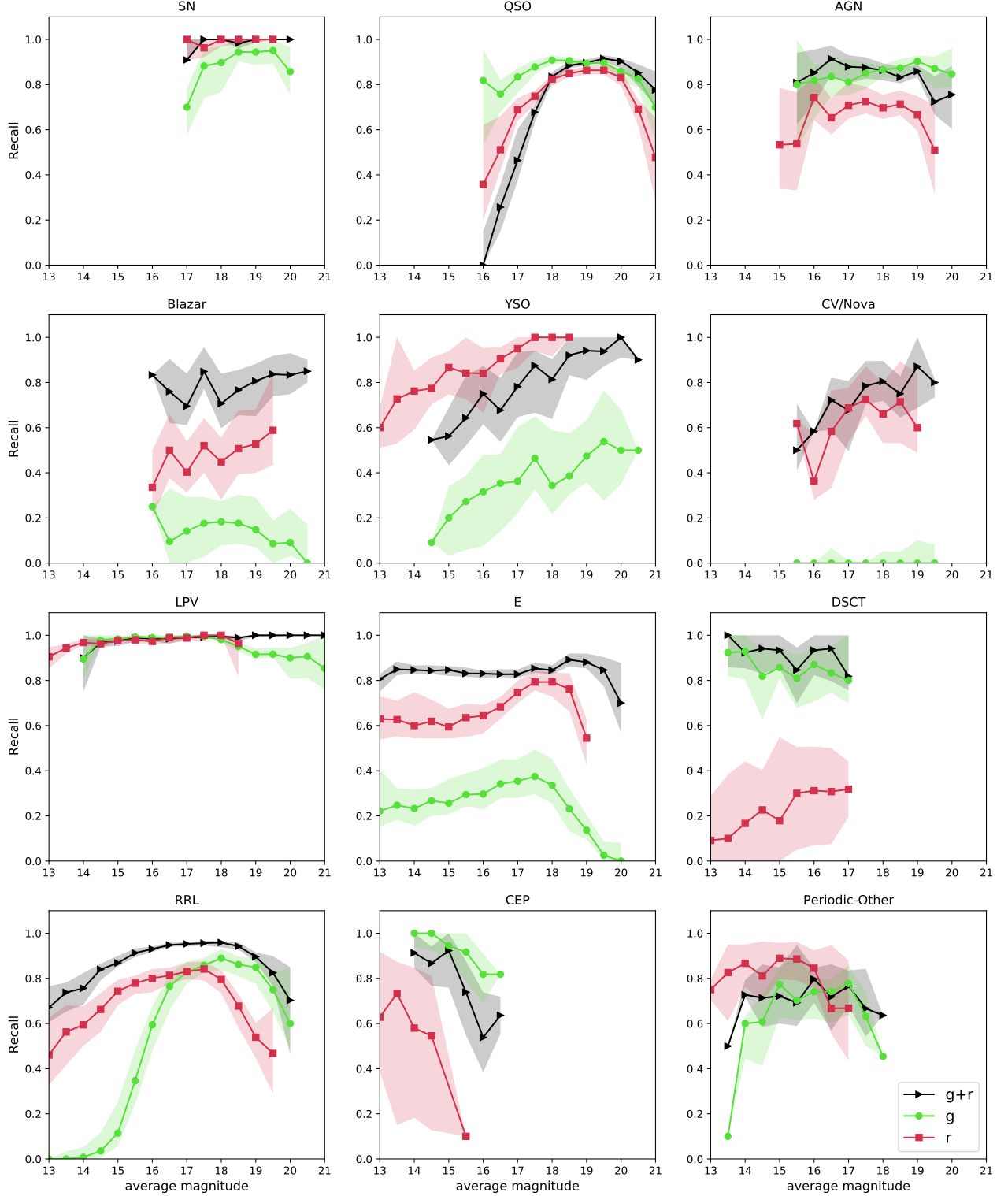
Figure 10. Recall for each stochastic and periodic subclass, as well as all transients (grouped as SN), as a function of the average magnitude. The x-axis ranges from 13 to 21 magnitudes, this range includes ∼90% of the sources. In black triangles we show the recall curves obtained when g and r photometries are available (considering the average magnitude in the g band), in green circles when only the g band is available, and in red squares when only the r band is available. The shaded regions were obtained by generating 20 different training and testing sets, and training 20 independent models using each of these sets. We report the median and 5 and 95 percentile values obtained from the 20 models. There is a truly wide variety of behaviors (see discussion in the text).

Figure 11. As in Figure 10, but plotting the Recall as a function of the number of detections in the light curve (in logarithmic scale). The x-axis ranges from 6 to 150 detections, this range includes ~90% of the sources. Again, there is a wide variety of behaviors (see discussion in the text).

True \ Pred	SNIa	SNIbc	SNII	SLSN	QSO	AGN	Blazar	YSO	CV/Nova	LPV	E	DSCT	RRL	CEP	Periodic-Other
SNIa	76	19	4	0	0	0	0	0	0	0	0	0	0	0	0
SNIbc	26	58	11	5	0	0	0	0	0	0	0	0	0	0	0
SNII	14	20	55	12	0	0	0	0	0	0	0	0	0	0	0
SLSN	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0
QSO	0	0	0	0	87	8	4	0	0	0	0	0	0	0	0
AGN	0	0	0	0	10	85	4	1	0	0	0	0	0	0	0
Blazar	0	0	0	0	14	9	75	0	1	1	0	0	0	0	0
YSO	0	0	0	0	0	0	1	78	0	5	1	0	1	4	11
CV/Nova	2	1	1	0	0	0	1	2	71	1	5	5	9	2	1
LPV	0	0	0	0	0	0	0	1	0	98	0	0	0	1	1
E	0	0	0	0	0	0	0	0	1	0	74	6	3	4	12
DSCT	0	0	0	0	0	0	0	0	0	0	6	89	2	1	2
RRL	0	0	0	0	0	0	0	0	0	0	4	4	88	2	2
CEP	0	0	0	0	0	0	1	2	0	1	3	1	9	75	8
Periodic-Other	0	0	0	0	0	0	0	2	1	1	8	3	6	6	73

Figure 12. As in Figure 8, but for the bottom level of the deployed BRF classifier.

`Meanvariance`, `ExcessVar`, and `SPM_A`), variability features that detect smooth decrease or increase of the luminosity (`LinearTrend`, `SPM_tau_rise`,`SPM_tau_fall`), features related with the quality of a supernova parametric model fitting (`SPM_chi`), and features related with transient appearance or disappearance (`positive_fraction`, and `n_non_det_after_fid`).

On the other hand, the Transient classifier is dominated by the SPM features (e.g., `SPM_gamma`, `SPM_beta`, `SPM_t0`, `SPM_tau_rise`, and `SPM_tau_fall`). Other relevant features are the optical colors in the peak and the mean of the light curve, measured from the difference image light curves, features that detect smooth increase or decrease of the observed flux (`LinearTrend`), features that measure the level of correlation in the light curve (`IAR_phi`), features related with the amplitude of the variability (`MHPS_low`), and features related with the appearance of a transient source (`dmag_first_det_fid`, `last_diffmaglim_before_fid_1`). Note that SN rise related features, such as `SPM_t0` and `SPM_rise`, are some of the most relevant features for the classification of transients, and are crucial for the early classification of SNe. Also, note that `SPM_t0` is not the explosion time, but some characteristic time where the SN has risen significantly.

For the Stochastic classifier, 12 of the top 30 features are related with color, morphology and distance to the Galactic plane, and the rest correspond to features related with the amplitude of the variabil-

ity observed at different time scales (e.g., `ExcessVar`, `SPM_A`,`Meanvariance`, `GP_DRW_sigma`, and `Amplitude`), and features related with the time scale of the variability (`IAR_phi`, `GP_DRW_tau`).

Finally, the Periodic classifier is clearly dominated by the `Multiband_period` feature, but also by different colors, by features related with the amplitude of the variability (e.g., `delta_mag_fid`, `Amplitude`, `ExcessVar`, `Meanvariance`, and `GP_DRW_sigma`), and features related with the timescale of the variations (e.g., `GP_DRW_tau`, and `IAR_phi`).

5.3.2. *Results for the unlabeled ZTF set*

We now turn to discuss the results obtained when applying the final BRF classifier to all the sources in ZTF with ≥ 6 detections in g or ≥ 6 detections in r. Considering the data obtained by ZTF until 2020/06/09, there are 868,371 sources that satisfy this condition, hereafter defined as the "unlabeled ZTF set". We define the class of a given object in the unlabeled ZTF set by selecting the class with the maximum probability obtained by the deployed BRF classifier. However, users of this classifier can use the obtained probabilities to make their own probability cuts and select samples for their science. The features, classifications and probabilities obtained for the unlabeled ZTF set with data until 2020/06/09 can be downloaded at Zenodo: 10.5281/zenodo.4279623.

It is important to note that the classifications obtained by the light curve classifier are updated every day, as new alerts are received. Whenever a new alert is received for a given source, the ALeRCE pipeline recomputes its variability features and provides an updated classification. These updated classifications can be found at the ALeRCE Explorer website, using the "light curve classifier" tab, and specifying the desired class. Considering the results shown in Figure 11, we expect that the quality of the classification for a given source will improve as more detections are added to the light curve. In addition, with new alerts, more objects will satisfy the condition of having ≥ 6 detections in g or ≥ 6 detections in r, and thus, the size of the unlabeled ZTF set increases every day. Moreover, with new detections the size of the labeled set will increase, allowing the training of new BRF models. Updates regarding any possible modification to the light curve classifier (e.g., labeled set and models) will be published on the ALeRCE Science website.

Figure 13 shows the number of candidates per class obtained for the 868,371 sources with enough alerts until 2020/06/09. Compared to Figure 4, it can be noticed that there is no correlation between the number of sources per class for the unlababel set and the number

Table 5. Feature ranking (top 30) for each layer of the deployed BRF classifier. Features marked with † correspond to non-detection features, features marked with ⋆ correspond to features computed using AllWISE data, and features marked with ‡ correspond to metadata features. Subscripts "_1" and "_2" refers respectively to g and r bands.

Top level		Transient		Stochastic		Periodic	
Feature	Rank	Feature	Rank	Feature	Rank	Feature	Rank
W1-W2⋆	0.094	SPM_t0_1	0.033	W1-W2⋆	0.109	Multiband_period	0.089
sgscore1‡	0.053	SPM_gamma_2	0.029	sgscore1‡	0.058	g-W2⋆	0.062
positive_fraction_2	0.050	SPM_tau_rise_2	0.028	r-W2⋆	0.049	r-W2⋆	0.034
positive_fraction_1	0.048	SPM_tau_rise_1	0.025	$(g-r)$_mean_corr	0.048	$(g-r)$_max_corr	0.030
SPM_tau_rise_1	0.035	$(g-r)$_max	0.023	g-W2⋆	0.046	g-W3⋆	0.028
LinearTrend_2	0.032	SPM_t0_2	0.022	gal_b‡	0.045	$(g-r)$_mean	0.027
SPM_chi_1	0.031	LinearTrend_2	0.019	g-W3⋆	0.037	$(g-r)$_max	0.025
g-W2⋆	0.031	AndersonDarling_2	0.018	$(g-r)$_max_corr	0.035	GP_DRW_tau_1	0.023
g-W3⋆	0.031	SPM_gamma_1	0.017	ExcessVar_2	0.033	$(g-r)$_mean_corr	0.022
n_non_det_after_fid_2†	0.026	SPM_tau_fall_2	0.017	Meanvariance_2	0.026	IAR_phi_1	0.022
W2-W3⋆	0.025	dmag_first_det_fid_1†	0.015	$(g-r)$_mean	0.025	Amplitude_1	0.017
SPM_gamma_1	0.024	MHPS_low_1	0.013	delta_mag_fid_2	0.024	ExcessVar_1	0.017
SPM_tau_rise_2	0.023	LinearTrend_1	0.013	r-W3⋆	0.023	delta_mag_fid_1	0.016
SPM_A_2	0.023	$(g-r)$_mean	0.012	W2-W3⋆	0.022	Meanvariance_1	0.016
SPM_chi_2	0.020	MHPS_ratio_1	0.011	Amplitude_2	0.022	r-W3⋆	0.016
SPM_A_1	0.019	SPM_tau_fall_1	0.011	Std_2	0.015	Std_1	0.015
r-W2⋆	0.018	SPM_beta_2	0.011	$(g-r)$_max	0.015	GP_DRW_sigma_1	0.015
ExcessVar_1	0.017	MHPS_ratio_2	0.010	SPM_A_2	0.014	GP_DRW_tau_2	0.012
ExcessVar_2	0.016	Skew_2	0.009	PercentAmplitude_2	0.013	PercentAmplitude_1	0.012
r-W3⋆	0.014	sgscore1‡	0.009	SPM_A_1	0.012	W1-W2⋆	0.009
Rcs_2	0.013	SPM_beta_1	0.009	ExcessVar_1	0.011	W2-W3⋆	0.009
GP_DRW_sigma_2	0.013	Power_rate_2	0.009	IAR_phi_1	0.010	SF_ML_amplitude_1	0.009
SPM_tau_fall_1	0.013	IAR_phi_2	0.009	GP_DRW_sigma_2	0.009	SPM_A_1	0.009
Meanvariance_2	0.012	dmag_first_det_fid_2†	0.009	delta_mag_fid_1	0.009	Gskew_1	0.009
LinearTrend_1	0.012	IAR_phi_1	0.009	Pvar_2	0.007	Q31_1	0.009
SPM_gamma_2	0.010	last_diffmaglim_before_fid_1†	0.009	IAR_phi_2	0.007	Autocor_length_1	0.009
Pvar_2	0.010	PPE	0.008	GP_DRW_tau_2	0.007	IAR_phi_2	0.008
MHPS_low_2	0.009	Harmonics_mag_6_1	0.008	GP_DRW_tau_1	0.007	SF_ML_gamma_1	0.008
SF_ML_amplitude_2	0.009	MHPS_low_2	0.008	MHPS_high_1	0.007	Amplitude_2	0.007
$(g-r)$_max_corr	0.009	Gskew_2	0.008	MHPS_low_1	0.006	delta_mag_fid_2	0.007

of sources per class for the labeled set. The Periodic-Other, E, and LPV classes have the highest number of candidates, while the SN classes have the lowest. The distribution of candidates per class is consistent with the astrophysical number densities (i.e., we are likely not misclassifying large numbers of sources). For instance, Blazars are relativistically beamed (and thus seen to have farther distances), but only over very small viewing angles, and hence are expected to be less common than QSOs and AGNs. In the case of SNe, not factoring in the amount of time a particular SN is above the magnitude limits of the search (the "control time"), we find ratios of SNe II/SNe Ia and SNe Ibc/SN Ia of 0.21 and 0.41, respectively. Computing these ratios using the number of such classes reported from ASAS-SN discoveries in Holoien et al. (2019) yields 0.36 and 0.09, respectively. The significant differences between the SNe Ibc/SN Ia

ratios implies that we are strongly overestimating the numbers of SN Ibc; given the similarities between SN Ibc and SN Ia light curves, we are likely classifying a non-negligible fraction of SN Ia as SN Ibc. This highlights the importance of including distance estimations to improve the classification of transients. We are currently working to include distance-based features in future versions of the classifier.

To investigate the quality of the predictions, we plotted the probability distributions of the top and bottom levels on the left and right sides of Figure 14. The red lines denote the position of the median probability for each class, and the green lines denote the 5 and 95 percentiles. It is clear from the figure that the distribution of probabilities for the top level are higher compared to the bottom level. For the top level, the classes with the lowest probabilities are CV/Nova, and YSO. For the
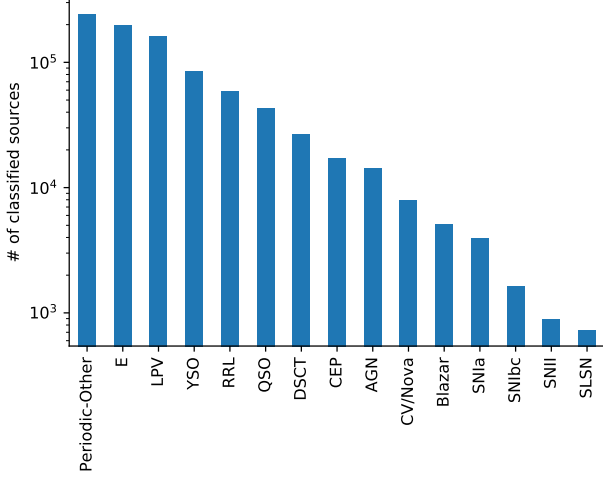
Figure 13. Number of candidates per class for all the sources in the unlabeled ZTF set. It can be seen that the number of sources per class for the unlabeled ZTF set does not correlate with the number of sources per class for the labeled set (see Figure 4).

bottom level, the classes with the highest probabilities are LPV, QSO, and AGN, and the lowest probabilities are for the different classes of SNe, YSO, CV/Nova, and some periodic variables.

The low probabilities obtained for some classes are related with the confusion between classes observed in Figure 7. For instance, in Figure 7 we can see that the SN classes present a high confusion among them. On the other hand, the SNIa, SNIbc, SNII and SLSN median probabilities of sources classified as SNII are 0.16, 0.19, 0.28 and 0.19, respectively. The high confusion among the SN classes may be due to the low number of sources in the labeled set, but also to the intrinsic similarities among these classes. For example, the physical mechanism responsible for the main peak of the light curve of SNe Ia and SNe Ib/c is the same, the diffusion of energy deposited by radioactive ^{56}Ni (Arnett 2008). Indeed, Villar et al. (2019a) report that 15% of their Type Ibc SNe are classified as Ia. This might be improved by performing data augmentation using, for example, Gaussian process modeling (e.g., Boone 2019). On the other hand, the low probabilities observed for CV/Novae and YSOs can be produced by the similarities between their colors and the colors of some periodic sources, and the fact that some CV/Novae and YSOs present very rapid variability compared to the ZTF cadence, that produces light curves with low auto-correlation, and thus low values of the `IAR_phi` parameter, which is normally observed for periodic sources (excluding LPVs). These similarities can be seen in Figure 15, where we show the distributions of `IAR_phi_1` and g-`W3` for YSOs and

CV/Novae (grouped), the rest of the stochastic classes, and periodic sources from the labeled set.

Figure 16 shows the normalized r band magnitude distribution of the different classes considering sources present in the labeled set and candidates from the unlabeled ZTF set. In general, the distributions of magnitudes of the candidates are similar to or fainter than found among sources from the labeled set. For instance, the SNe classes have candidates that are ~ 0.5 magnitudes fainter than the labeled set, and the YSO, CV/Nova and E classes have candidates that are ~ 1 magnitude fainter than the labeled set. These results show that the classifier is able to detect faint and bright candidates, regardless of the luminosity biases present in the labeled set, which can be dominated by the brightest tail of the true magnitude distribution of each class. This will be particularly relevant for surveys like LSST, since, in general, available training set will be $\sim 2 - 3$ magnitudes brighter than the limiting magnitudes of the single images, and thus we would expect that currently existent bright samples will allow us to detect fainter candidates.

A simple way to test the performance of the BRF classifier is to verify whether the results obtained when the model is applied to the unlabeled ZTF set are in agreement with what is astrophysically expected from previous works. For instance, younger Galactic targets like YSOs, Classical Cepheids, and LPVs should reside near the Galactic plane (e.g., Catelan & Smith 2015; Mowlavi et al. 2018), while extragalactic sources like AGNs, QSOs, Blazars, and SNe should have roughly isotropic distributions, perhaps with fewer sources near the Galactic plane due to attenuation/reddening by gas and dust (e.g., Calzetti et al. 2000; Padovani et al. 2017). On the left side of Figure 17 the sky distribution, in Galactic coordinates, of LPV, CEP, and YSO candidates is shown. It is clear from the figure that most of them are located in the Galactic plane, and that sources located outside the plane have a low BRF probability. This is consistent with the results obtained by previous works (e.g., Mowlavi et al. 2018; Rimoldini et al. 2019). The right panel of Figure 17 shows the Galactic latitude versus the $g - r$ color obtained using the mean magnitude of the light curves in each band, for extragalactic candidates (QSO, AGN, Blazar, SNIa, SNIbc, SNII, and SLSN). From the figure we can see that the fraction of extragalactic candidates observed around the Galactic plane is low, and that most of the candidates located in the plane have low probabilities. Moreover, the $g - r$ colors of the extragalactic candidates are consistent with what is expected for these classes, with clear evidence of reddening for the candidates located around the Galac-
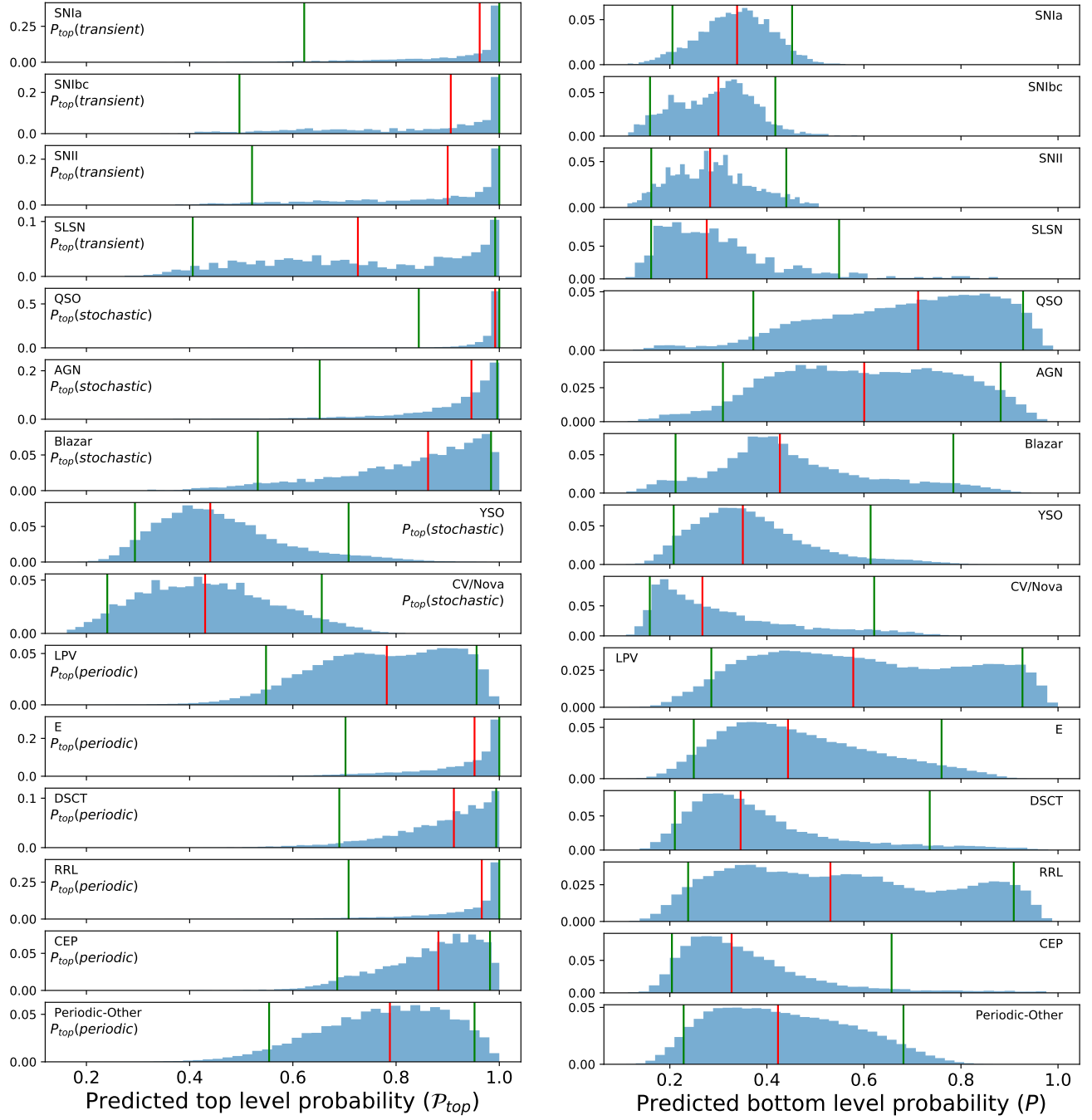
Figure 14. Left: normalized probability distributions of the top level of the deployed BRF classifier, split by subclass, for candidates from the unlabeled ZTF set. The reported values correspond to the probabilities obtained for each class of the top level, as indicated below the class name. Right: normalized probability distributions of the bottom level of the deployed BRF classifier, split by subclass, for candidates from the unlabeled ZTF set. The red lines show the median probability for each class. The green lines show the 5 and 95 percentiles of the probabilities. Some subclasses show broad distributions to low values, implying that they are not so well-represented or characterized by the highest ranked features within the hierarchical class.

Figure 15. Normalized IAR ϕ distribution (*top*) in the g band (`IAR_phi_1`; cut in the y-axes at 0.4), and normalized g-W3 distribution (*bottom*), for YSOs and CV/Novae (blue), stochastic sources (red; excluding CV/Novae and YSOs), and periodic sources (yellow). It can be seen that there is an overlap in the `IAR_phi_1` and g-W3 distributions of CV/Novae, YSOs, and periodic sources.

tic plane. The sky distribution, in Galactic coordinates, of the extragalactic candidates can be found in Figure 24 of the appendix.

6. CONCLUSIONS

6.1. *Summary*

In this paper we presented the first version of the ALeRCE light curve classifier. This classifier uses a total of 152 features, including variability features computed from ZTF light curves with ≥ 6 epochs in g or r bands, and colors computed using ZTF and AllWISE photometry (see Section 3), to classify each source into

Figure 16. Normalized magnitude distributions in the r band for sources in the labeled set (LS; red histograms) and candidates from the unlabeled ZTF set (cand.; blue histograms).

15 subclasses, including periodic, transient, and stochastic variable sources (see Section 2.2). The light curve classifier uses a balanced RF classifier (see Section 4.1), constructed with a two-level scheme. The first level (top level) classifies each source as periodic, stochastic or transient. The second level (bottom level) consists

Figure 17. Left: Galactic latitude (gal_b, in degrees) versus Galactic longitude (gal_l, in degrees) for candidates expected to be mostly observed around the Galactic plane (LPV, CEP, and YSO classes). Right: Galactic latitude (gal_b, in degrees) versus $g - r_{max}$ for extragalactic candidates (QSO, AGN, Blazar, SNIa, SNIbc, SNII, and SLSN classes). The bottom level probability computed by the deployed BRF classifier are color-coded according to the colorbar to the right. The red contours show the density of points in each plot. It can be seen that most of the extragalactic candidates are located outside the Galactic plane, while most of the YSO, LPV and CEPcandidates are located in the Galactic plane.

of three classifiers that further resolve each hierarchical class into subclasses.

We trained and tested the BRF classifier using a labeled set obtained by cross-matching the ZTF database with different catalogs of transients, stochastic and periodic sources (see Section 2.2.1). For the top level we obtained macro-averaged precision, recall, and F1-score values of 0.96, 0.99, and 0.97, respectively, while for the bottom level we obtained macro-averaged precision, recall, and F1-score values of 0.57, 0.76, and 0.59, respectively.

We used the BRF classifier to classify 868,371 sources from ZTF (unlabeled ZTF set), obtaining results that are in agreement with what we expect astrophysically. For instance, most of the high probability extragalactic candidates are located outside the Galactic plane, and most of the high probability YSO, LPV, and CEP candidates are located in the Galactic plane.

The condition of ≥ 6 detections in g or r normally equates to a timespan between 3 and 30 days since the first detection. Whenever a new detection is received for an object, the ALeRCE pipeline processes it and provides an updated classification in ~ 1 second. The light curve classifier provides updated classifications for objects with new ZTF alerts every day. These updated classifications can be found on the ALeRCE Explorer website, selecting the "Light curve

classifier" option, and specifying the desired class. Catalogs containing the labeled set, and the features and RF probabilities obtained by the top and bottom levels for the unlabeled ZTF set (up to 2020/06/09) can be downloaded at Zenodo: 10.5281/zenodo.4279623. In addition, more examples and instructions on how to use the ALeRCE database and classifications can be found on the ALeRCE Science website and in Förster et al. (2020), and a detailed description of the ALeRCE database can be found in the database schema. Finally, the code used to train the deployed BRF classifier can be found on the "light_curve_classifier_SanchezSaez_2020" GitHub repository, and the implementation of this classifier in the ALeRCE pipeline can be found on the "lc_classifier" GitHub repository, which version 1.0.1 is archived in Zenodo (Sánchez-Sáez et al. 2020).

6.2. *Final remarks and perspectives*

One of the main challenges found during the development of the ALeRCE light curve classifier was the high imbalance present in the labeled set. For instance, the transient sources represent 1.4% of the labeled set, while the periodic sources represent 70.5%. Each hierarchical class also suffers from high imbalance among its subclasses; for example, in the case of the transient class, SNIa comprise 74.0% of the sample, while SLSN correspond to only 1.4%. We addressed this prob-

lem by using the balanced RF implementation of the `imbalanced-learn` Python package, which follows the procedure proposed by Chen et al. (2004). This method uses a downsampling majority class technique to train each tree with a balanced sub-sample. We also tested two other algorithms, GBoost and MLP, but concluded that more work is needed if we want to obtain better results from those.

Another challenge was to find features useful to separate the different classes. Previous works have normally used features similar to those available in the FATS Python package (e.g., Kim et al. 2014; Martínez-Palomera et al. 2018), however our first tests demonstrated that these features were not informative enough to separate the 15 classes considered by the light curve classifier, in particular the stochastic and transient classes. Thus, novel features were designed and implemented for this work, like the `IAR_phi` parameter, the MHPS features, and the non-detection features. In addition, during the development of the light curve classifier we realized that some stochastic classes are hard to separate using just variability features. In particular, the separation of YSOs from the other stochastic classes improved significantly once we included AllWISE colors in the set of features.

Furthermore, the computation of reliable periods was quite challenging, particularly considering that the ALeRCE pipeline requires fast computation of features. Huijse et al. (2018) demonstrated that very good results can be achieved by quadratic mutual information (QMI) estimators, however these techniques are computational expensive $[O(n^2)]$. We solve this issue by using the MHAOV periodogram, which provides less reliable periods, but is much faster to compute $[O(n)]$. Periods become increasingly unreliable as the number of datapoints decreases, but the classification of periodic variables can still be accurate, as other features can compensate for the decreasing quality of the periodogram (e.g., features related with the amplitude and the timescale of the variability). ALeRCE is currently working to implement methods for period estimation that are both accurate and fast. Computing the periodogram is expensive for sources with a large number of detections. We are currently exploring so-called "online" periodograms, which are updated as new samples arrive, at a fraction of the computational cost of recomputing the period each time from scratch (Zorich et al. 2020), as well as other techniques that might work better with eclipsing binary light curves (e.g., Kovács et al. 2002; Mighell & Plavchan 2013).

Moreover, the classification of the different SN classes was particularly challenging. First, the number of SNe in the labeled set was very small compared to other classes, and second, the light curves of SN classes can present similarities, which makes their separation difficult, as discussed in Section 5.3.2. We solved this issue by using the `BalancedRandomForestClassifier` method from `imbalanced-learn`, and by including the SPM features, whose definition is a modification of the work of Villar et al. (2019a). In the future we plan to test other techniques to improve the separation of SN classes. Previous works have performed Gaussian process regression to model SNe light curves and generate new light curves with different cadences therefrom (e.g., Boone 2019). Moreover, better results can be obtained if we use information regarding the SN host galaxy (e.g. Foley & Mandel 2013; Baldeschi et al. 2020).

In the future we also plan to perform data augmentation to improve the classification of variable objects. For the case of variable stars, light curves can be modeled with Gaussian process or with a combination of harmonics, and then basic transformations can be applied to these models to obtain light curves with different periods and amplitudes (e.g., Elorrieta et al. 2016; Martínez-Palomera et al. 2018; Castro et al. 2018; Aguirre et al. 2019; Hosenie et al. 2020). To the best of our knowledge for the case of AGNs, QSOs and Blazars no previous attempts to perform data augmentation have been made. A promising option is to use synthetic light curve generators that consider the physical processes behind the variability (e.g., Sartori et al. 2019).

Most of the features used by this classifier can be implemented and used to classify light curves from other data sets. In particular, for the case of LSST, the non-detection features can be adapted to work with the forced photometry that will be provided for each alert (`DIAForcedSources` in the Data Products Definition Document; Jurić et al. 2019). LSST will also benefit from the multiband *ugrizy* light curves. As we demonstrated in this work, in general the light curve classification improves when both ZTF *g* and *r* data are available. For the case of LSST this would be the same, and probably we should even be able to further resolve some of the subclasses presented in this work. For instance, using the *zy* light curves we should be able to separate local type 1 and type 2 active galactic nuclei, since for low redshift sources, we can detect variability from the dusty torus at these wavelengths (see Sánchez et al. 2017 and references therein), or identify high redshift QSOs, whose emission is expected to be absorbed in the bluer bands. We encourage researches interested in classifying stochastic and transient sources in particular to use the novel (or modified) features presented in this work, like

the `IAR_phi` parameter, the MHPS features, the SPM features, and the non-detection features.

It is worth to note that the ALeRCE light curve classifier is being constantly improved, and this work describes its Version 1.0. Future versions of this classifier may include new classes of variable and transient objects, as well as sub-classes of sources already present in the taxonomy (e.g., RRL types ab and c; classical and type II Cepheids; contact, detached, and semi-detached eclipsing binaries; among others). We are also working to find new features and techniques that can improve the performance of the classifier. Future work will report any changes included in the classifier model, like the inclusion of data augmentation, or the use of other classification strategies (e.g., semi supervised training). We recommend the users of this classifier to check the ALeRCE Science website to get updates related with the different classifiers and the data processing. We are exploring different classification algorithms which are not based on manually designed features, but on automatically derived, recurrent, implicitly extracted features, via deep learning (e.g. Naul et al. 2018; Muthukrishna et al. 2019; Becker et al. 2020). However, up to this point we have found that the former produce better results when applied to real data. Most likely, a combination of simulated and real data will be required to train reliable deep learning classification models in the future, as found by Carrasco-Davis et al. (2019).

ACKNOWLEDGMENTS

APPENDIX

---

[5] Python and R implementations are available in https://github.com/felipeelorrieta/IAR_Model

A. FURTHER DESCRIPTION OF SOME VARIABILITY FEATURES

In this section we provide additional description of some of the features listed in Table 2 (those marked with **). These features correspond to new variants of features included in the FATS package and other works:

- Damp Random Walk (DRW) parameters: a DRW model is defined by a stochastic differential equation which includes a damping term that pushes the signal back to its mean: $dX(t) = -\frac{1}{\tau_{DRW}}X(t)dt + \sigma_{DRW}\sqrt{dt}\,\epsilon(t) + b\,dt$, $\tau_{DRW}, \sigma_{DRW}, t > 0$. τ_{DRW} corresponds to the characteristic time for the time series to become roughly uncorrelated, σ_{DRW} corresponds to the amplitude of the variability at short timescales ($t \ll \tau_{DRW}$), and $\epsilon(t)$ is a white noise process with zero mean and variance equal to 1. DRW modelling is typically used to describe light curves of active galactic nuclei (Kelly et al. 2009). In this case we obtained the σ_{DRW} and τ_{DRW} parameters using Gaussian process regression, with a Ornstein-Uhlenbeck kernel, as in Graham et al. (2017). `GP_DRW_sigma` denotes σ_{DRW}, while `GP_DRW_tau` denotes τ_{DRW}.

- Excess Variance ($\sigma_{\rm rms}$): Measure of the intrinsic variability amplitude in a given band (see Sánchez et al. 2017, and references therein). $\sigma_{\rm rms}^2 = (\sigma_{LC}^2 - \overline{\sigma}_m^2)/\overline{m}^2$, where σ_{LC} is the standard deviation of the light curve, $\overline{\sigma}_m$ is the average photometric error, and $\overline{m}$ is the average magnitude. We denoted $\sigma_{\rm rms}^2$ as `ExcessVar`.

- Multiband Period: The period is estimated using the Multi Harmonic Analysis of Variance (MHAOV) periodogram (Schwarzenberg-Czerny 1996), which is based on fitting periodic orthogonal polynomials to the data. A single period estimate per light curve is computed by fitting both bands using the MHAOV multiband extension proposed by Mondrik et al. (2015). We denote this period as `Multiband_period`. For sources with detections only in g or only in r, the `Multiband_period` reports the single band period. To avoid overfitting the data when few samples are available we set the number of harmonics to one. This might not capture the best period for non-sinusoidal light curves, e.g., detached and semi-detached eclipsing binaries, returning a harmonic instead. We found that having a harmonic of the true period is in general sufficient to classify non-sinusoidal light curves correctly given that other features such as the `power rate` are included. We choose MHAOV for this analysis as it provides a good trade-off between performance and computational complexity. This method is now implemented and available in the `P4J` package (Huijse et al. 2018).

- Harmonics parameters: Harmonic series (Stellingwerf & Donohoe 1986) are commonly used to model and classify periodic light curves (Debosscher et al. 2007; Sarro et al. 2009; Richards et al. 2011; Elorrieta et al. 2016). In this work we fit a harmonic series up to the seventh harmonic, according to the expression

$$y(t_j) = \sum_{k=1}^{7}\left[A_k \cos\left(\frac{2\pi k t_j}{P}\right) + B_k \sin\left(\frac{2\pi k t_j}{P}\right)\right] + C, \tag{A1}$$

where t_j corresponds to the observational time of the j-th detection, P is the best candidate period computed from the multiband periodogram as above, and $y(t_j)$ is the magnitude estimated by the harmonic model. Even though we use the `Multiband_period`, the harmonic model is computed using the detections in each band independently. A_k, B_k and C for $k = 1, \ldots, 7$ are obtained by minimizing the weighted mean square error between the observed magnitudes and the model.

Note that the model is linear with respect to its parameters, so the latter can be computed using weighted linear regression. The inverse of the square of each observational error is used as a weight, which minimizes contributions from noisier observations. The cost function is given by

$$\min_{A_k, B_k} \frac{1}{J}\sum_{j=1}^{J}\frac{[\mathrm{mag}(t_j) - y(t_j)]^2}{\mathrm{sigma}(t_j)^2}, \tag{A2}$$

where J is the number of observations, $\mathrm{mag}(t_j)$ is the observed magnitude at time t_j, and $\mathrm{sigma}(t_j)$ is the observational error at time t_j. The solution to the weighted least squares optimization problem is found using the Moore-Penrose pseudoinverse. This solution has the additional property of having the minimum Euclidean norm when the problem is underdetermined (Ben-Israel & Greville 2003), which in this case corresponds to having less than 15 observations.

Once the parameters are learnt, equation A1 can be rewritten as

$$y(t_j) = \sum_{k=1}^{7} M_k \cos\left(\frac{2\pi k t_j}{P} - \phi_k\right) + C, \tag{A3}$$

with $M_k = \sqrt{A_k^2 + B_k^2}$ and $\phi_k = \arctan(B_k/A_k)$. In this way, the harmonics are now described by the amplitude and phase of each component. The model is shifted in time in order to have zero phase in the first harmonic, which is done following the expression $\phi'_k = \phi_k - k\phi_1$, replacing ϕ_k by ϕ'_k in Eq. A3.

Finally, the parameters M_k for $k = 1, \ldots, 7$, ϕ'_k for $i = 2, \ldots, 7$ and the mean square error are used as features, which are denoted `Harmonics_mag_1`, ..., `Harmonics_mag_7`, `Harmonics_phase_2`, ..., `Harmonics_phase_7`, `Harmonics_mse`, respectively.

- P_{var}: Probability that the source is intrinsically variable in a given band (see Paolillo et al. 2004, and references therein). It considers the χ^2 of the light curve respect to its mean, and calculates the probability $P_{var} = P(\chi^2)$ that a χ^2 lower or equal to the observed value could occur by chance for an intrinsically non-variable source, assuming that for each light curve its χ^2 will follow a probability distribution described by an incomplete gamma function $\Gamma(\nu/2, \chi^2/2)$, where ν corresponds to the degrees of freedom. We denoted P_{var} as `Pvar`.

- Structure Function (SF) parameters: The SF quantifies the amplitude of the variability as a function of the time difference between pairs of detections (τ). In this work we consider the definition provided by Caplar et al. (2017). We model the SF as a power law: $\text{SF}(\tau) = A_{\text{SF}}\left(\frac{\tau}{1\text{yr}}\right)^{\gamma_{\text{SF}}}$, where γ_{SF} corresponds to the logarithmic gradient of the change in magnitude, and A_{SF} corresponds to the amplitude of the variability at 1 yr. `SF_ML_amplitude` denotes A_{SF}, while `SF_ML_gamma` denotes γ_{SF}.

- Supernova parametric model (SPM): Villar et al. (2019b) introduced an analytic model describing SN light curves as a six parameter function, which they used to characterize and classify SN light curves from the Pan-STARRS1 Medium-deep Survey (Chambers et al. 2016). This model is an extension of previous empirical efforts to analytically describe supernova light curves, including the effects of different explosion times, normalization factors, initial rise timescales, rate of decline after peak, plateau lengths, or tail decay timescales. We introduce two modifications to this model. First we reparametrize the function to always remain positive in a simple validity range by a set of inequalities. After the first modification, the model is the following:

$$F = \begin{cases} \dfrac{A\left(1 - \beta'\frac{t-t_0}{t_1-t_0}\right)}{1 + \exp\left(-\frac{t-t_0}{\tau_{\text{rise}}}\right)} & \text{if } t < t_1 \\[20pt] \dfrac{A(1-\beta')\exp\left(-\frac{t-t_1}{\tau_{\text{fall}}}\right)}{1 + \exp\left(-\frac{t-t_0}{\tau_{\text{rise}}}\right)} & \text{if } t \geq t_1, \end{cases} \tag{A4}$$

where we also use $\gamma \equiv t_1 - t_0$ as a parameter instead of t_1. This function is positive valued when $A > 0$, $\gamma > 0$, $\tau_{\text{rise}} > 0$, $\tau_{\text{fall}} > 0$ and $0 < \beta' < 1$.

The second difference with respect to Villar et al. (2019a) is replacing the piecewise-defined function for a soft transition between the two components. This is done by including a sigmoid function $\sigma(t) = 1/(1 + \exp(-t))$, which allows a soft transition between zero and one. As the parameter t_1 defines the transition between the two pieces of the model in Eq. A4, it cannot be optimized properly using first-order methods. Our proposed model allows using this technique effectively to learn the parameters of the model, which is given by the following equation:

	SNIa	SNIbc	SNII	SLSN	QSO	AGN	Blazar	YSO	CV/Nova	LPV	E	DSCT	RRL	CEP	Periodic-Other
SNIa	69	22	4	1	0	2	0	0	0	0	0	0	0	0	1
SNIbc	39	29	11	11	0	7	0	4	0	0	0	0	0	0	0
SNII	14	21	51	10	0	1	0	0	0	0	0	0	0	0	2
SLSN	29	0	0	71	0	0	0	0	0	0	0	0	0	0	0
QSO	0	0	0	0	86	7	6	1	0	0	0	0	0	0	0
AGN	0	0	0	0	16	73	9	2	0	0	0	0	0	0	0
Blazar	0	1	0	0	21	16	56	4	1	1	0	0	0	0	0
YSO	0	0	0	0	0	3	2	79	2	4	4	0	1	2	3
CV/Nova	3	3	1	0	0	0	2	1	70	1	3	6	6	1	3
LPV	0	0	0	0	0	0	0	8	0	90	0	0	0	1	1
E	0	0	0	0	0	0	0	2	2	0	65	9	3	5	13
DSCT	0	0	0	0	0	0	0	0	0	0	5	77	5	3	10
RRL	0	0	0	0	0	0	0	1	0	0	3	12	71	2	11
CEP	0	0	0	0	0	0	1	2	1	3	6	3	14	59	11
Periodic-Other	0	0	0	0	0	0	1	2	5	1	9	6	3	1	71

Predicted label

Figure 18. Confusion matrix for a one-level multi-class RF Model. The black squares highlight the three hierarchical classes (from top to bottom, transient, stochastic, and periodic, respectively). To normalize the confusion matrix results as percentages, we divide each row by the total number of objects per class with known labels. We round this percentages to integer values. The performance of this model is poorer compared to the BRF classifier (see Figure 7).

$$F = \frac{A\left(1 - \beta'\frac{t-t_0}{t_1-t_0}\right)}{1 + \exp\left(-\frac{t-t_0}{\tau_{\text{rise}}}\right)} \cdot \left[1 - \sigma\left(\frac{t-t_1}{3}\right)\right] + \frac{A(1-\beta')\exp\left(-\frac{t-t_1}{\tau_{\text{fall}}}\right)}{1 + \exp\left(-\frac{t-t_0}{\tau_{\text{rise}}}\right)} \cdot \left[\sigma\left(\frac{t-t_1}{3}\right)\right]. \tag{A5}$$

In this particular model, for all the sources we use the light curves based on the difference images (`lc_diff`). This is done to avoid the contamination from unrelated host galaxy emission, which can distort the real shape of the SNe light curves. We also subtract from t the MJD value of the first detection observed for a given source. We computed A (`SPM_A`), β' (`SPM_beta`), t_0 (`SPM_t0`), γ (`SPM_gamma`), τ_{rise} (`SPM_tau_rise`), and τ_{fall} (`SPM_tau_fall`), for each band independently. In addition, we computed the reduced χ^2 of the fit for the light curve, denoted as `SPM_chi`. The parameters are found using the function `curve_fit` provided by the Scipy library (Virtanen et al. 2020).

B. ONE-LEVEL MULTI-CLASS RF MODEL

The first model tested for the ALeRCE light curve classifier was a simple one-level RF model with 15 classes, implemented using the `imbalanced-learn` Python package. This model uses 500 trees, maximum depth trees, and maximum number of features equal to the square root of the total number of features. Figure 18 shows the confusion matrix obtained by this model. The precision, recall and F1-score obtained are 0.49, 0.68, and 0.50, respectively. Clearly this model has a lower performance compared to the BRF classifier.

C. FURTHER ANALYSIS OF CLASSES WITH ANOMALOUS RECALL CURVES

C.1. *The particular case of AGN, QSO and Blazar*

In this work we present the first attempt to separate different types of active galactic nuclei according to their variability properties. As we mentioned in Section 2.2, we separate active galactic nuclei in the following way:

- AGN: type 1 Seyfert galaxies (i.e., active galactic nuclei whose emission is dominated by the host galaxy), selected from MILLIQUAS (broad type "A"), and from Oh et al. (2015).

- QSO: type 1 core-dominated active galactic nuclei (i.e., active galactic nuclei whose emission is dominated by their active nuclei), selected from MILLIQUAS (broad type "Q").

- Blazar: BL Lac objects and Flat Spectrum Radio Quasars (FSRQ), selected from ROMABZCAT and MILLI-QUAS.
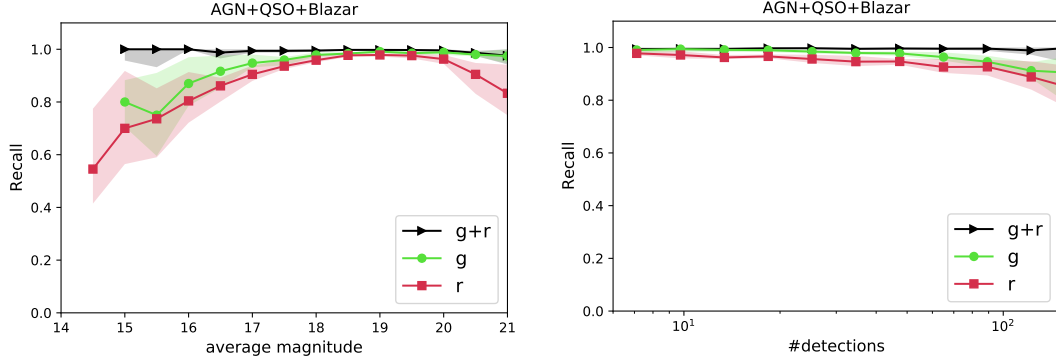

Figure 19. Similar to Figures 10 and 11, but treating AGNs, QSOs, and Blazars as a single class. When grouped together, the recall is much better behaved than for any individual active galactic nuclei class, highlighting remaining difficulties distinguishing between these classes.

In Figure 11 we showed the recall curves for each class as a function of the number of detections, and we could observe that these curves decrease for QSOs and AGNs when more detections are available, particularly in the r band. These results are puzzling, considering that we would normally expect to improve the classification of a given class as more detections are included in the light curves. In order to better understand the origin of these results, we show in Figure 19 the recall curves as a function of average magnitude and number of detections for AGNs, QSOs and Blazars grouped as a single class. In this case, the recall curves are around 0.8 and 1.0 for every bin of magnitude (specially for $g > 16$ or $r > 16$) and number of detections. From this we can infer that the light curve classifier has a very good performance selecting active galactic nuclei as a single class, but some issues still remain regarding the separation of AGNs, QSOs, and Blazars. There are two main explanations for these results: a) the method cannot properly separate QSOs from AGNs and Blazars, or b) there are sources in the labeled set with incorrect labels.

A possible way to explore how well the light curve classifier can discriminate among AGNs, QSOs, and Blazars, is to check whether the features available in the light curve classifier can separate these three populations. Figure 20 shows six different features used by the classifier, $(g-r)$_mean_corr, Meanvariance_1, ExcessVar_1, sgscore1, Meanvariance_2, and ExcessVar_2, for QSOs, AGNs, and Blazars from the labeled set (most of these features are in the top 30-ranked features shown in Table 5). From the figure we can see that these three classes have different color distributions, different morphologies, and also different variability properties. AGNs and Blazars tend to be redder than QSOs (see $(g-r)$_mean_corr), Blazars and AGNs tend to have larger amplitudes (see Meanvariance_1 and ExcessVar_1), and AGNs tend to have more extended morphologies compared to QSOs and Blazars. These are just some examples of features that can be used to separate the three classes above mentioned. After a visual inspection of the feature distribution of AGNs, QSOs, and Blazars, we found that more than 30 features can be used to separate them, including for instance PercentileAmplitude, Q31, GP_DRW_sigma, GP_DRW_tau, MHPS_low, MHPS_high, SF_ML_amplitude, among others. From this we can infer that the light curve classifier should be able to separate these three populations.

In addition, Figure 20 highlights that the g band features seem to separate better the AGN and QSO classes compared to same features in the r band. This behavior is also seen in other features, like PercentileAmplitude, GP_DRW_sigma, Std, among other features related with the amplitude of the variability. These differences might be produced by the combined effect of having a higher contamination from the host in the r band and intrinsically lower amplitude of the variability in the r band, due to the well known anti-correlation between amplitude of the variability and the wavelength of emission (see Sánchez et al. 2017 and references therein). From this, we can understand the differences observed in the g and r band recall curves of AGNs and QSOs shown in Figure 11, which should be produced by these differences in the features distributions.
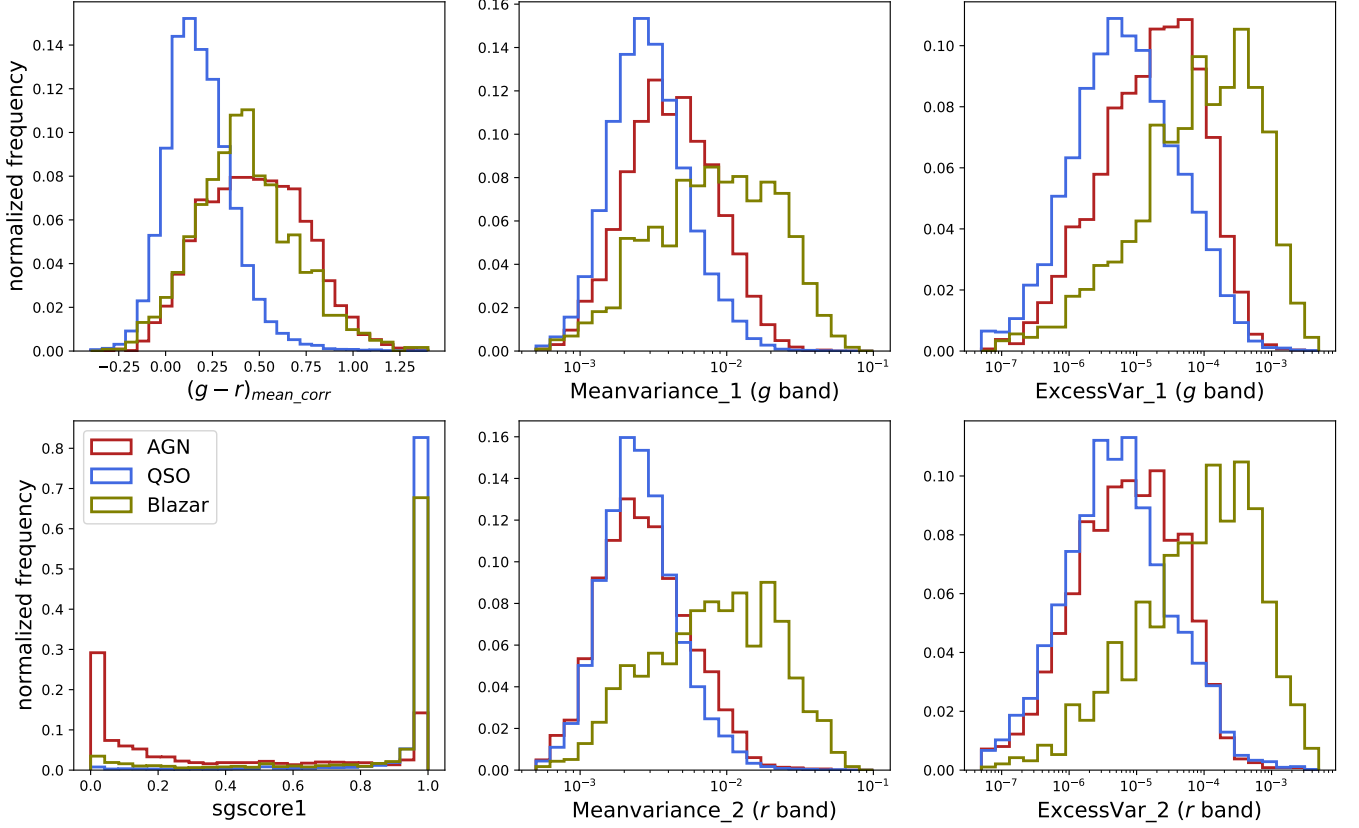
Figure 20. Distribution of $(g-r)$_mean_corr, Meanvariance_1, ExcessVar_1, sgscore1, Meanvariance_2, and ExcessVar_2, for QSOs (blue), AGNs (red), and Blazars (yellow) classes from the labeled set. These features (among others) can be used to separate these classes of active galactic nuclei. Subscripts "_1" and "_2" refers respectively to g and r bands.

On the other hand, one key source of confusion between AGN and QSO is the definition of this division criteria in the labeled set. MILLIQUAS uses a luminosity-based division to separate AGNs and QSOs (see Section 5 of Flesch 2015), while the variability amplitude features are also affected by the ratio between core and host luminosities. Our variability-based criteria appears to separate host-dominated from core-dominated sources, in rough correspondence with low and high core luminosities, respectively. In addition, it is hard to make strict cuts to separate the AGN, QSO, and Blazar classes, considering their observational properties (e.g., luminosity, redshift, orientation). In particular, the dividing line between AGN and QSO has never been defined properly (e.g., Netzer 2013; Padovani et al. 2017), and thus different catalogs will provide different classifications for the same object. Note that since the redshift of the sources in the unlabeled set is generally unknown, luminosity cannot be used as a feature by the classifier. Our aim is to separate host-dominated from core-dominated sources to favor the identification of different populations of active galactic nuclei by the light curve classifier, and use the AGN and QSO classifications as representative of these classes in the extremes of the luminosity distribution, but expect that there will be strong mixing of these classes close to the dividing luminosity used in the labeled set.

In order to see whether there are sources in the labeled set with incorrect QSO or AGN labels, we crossmatched our labeled set with the SIMBAD database. There are 26168 QSOs in the labeled set (all obtained from MILLIQUAS), and 1590 of them are classified as Seyfert or AGN by SIMBAD (6%), with 1580 having a reported redshift < 1. On the other hand, there are 4667 AGNs in the labeled set, and 830 (17%) of them are classified as QSO in SIMBAD. Therefore, there are 2420 sources in the labeled set with inconsistent classification in MILLIQUAS and SIMBAD. The light curve classifier classifies 920 of these sources as QSO and 1319 as AGN.

To understand better this discrepancy in the classification of some sources, we crossmatched our labeled set with the catalog of spectral properties of Quasars from SDSS DR14 provided by Rakshit et al. (2020), as well as with the catalog of Oh et al. (2015), in order to obtain bolometric luminosities (L_{bol}), BH masses (BH_{mass}), Eddington ratios

(L/L$_{\mathrm{Edd}}$) and redshifts, for our sample of QSOs and AGNs. We excluded Blazars from this analysis, since they are not properly identified and characterized in these catalogs.

We show in Figure 21 the distribution of these spectral properties for AGNs and QSOs from the labeled set and for misclassified sources (i.e., AGNs from the labeled set classified as QSO, and vice versa). It is clear from the figure that AGNs and QSOs from the labeled set have spectral property distributions that are not strictly separated. From this we can conclude that the luminosity-based division performed by MILLIQUAS is not physically so meaningful. One issue, for instance, is to what extent the host contributes to the total magnitude used to classify the source as AGN or QSO. In addition, Figure 21 shows that the misclasified sources have similar distributions of redshift and luminosity. Moreover, these distributions lie exactly in the range of values where the properties of AGNs and QSOs from the labeled set overlap (redshift$\sim$ 0.5 and L$_{\mathrm{bol}} \sim 10^{45.5}$ [erg/s]).

Figure 21. Distribution of redshift, BH$_{\mathrm{mass}}$, L$_{\mathrm{bol}}$, and L/L$_{\mathrm{Edd}}$ of AGNs (red dashed line) and QSOs (blue dashed line) from the labeled set, and missclassified AGN (green solid line) and QSO (yellow solid line) candidates. The spectral properties where obtained from catalogs that make use of SDSS spectra (Oh et al. 2015; Rakshit et al. 2020)

From these results, we can infer that the decrease in the recall as a function of the number of detections observed for QSOs and AGNs is produced by the discrepancy in the QSO/AGN classification obtained from different catalogs, which is produced by the difficulty of making strict cuts to separate AGNs and QSOs. When more epochs are available, it is easier for the light curve classifier to perform a correct variability-based classification, and therefore, identify an original inconsistent classification provided by a given catalog. We propose that by using the variability properties of active galactic nuclei we can more easily separate them as core-dominated or host-dominated, or in other words, as bright QSOs or low luminosity AGNs.

C.2. *The particular case of RRL*

In Section 5.2 we claimed that the low recall values obtained when only the g band photometry is used for bright RRL can be explained by the differences in the variability features of the RRL sub-types. Figure 22 shows the

`Multiband period` versus the `Meanvariance` measured in the g and r bands, for bright RRLs (mean $g < 16$) split into their sub-classes 'ab' and 'c', and for E, CEP, and DSCT (grouped as a single class). From the figure we can notice that ab-type RRL tend to have larger `Meanvariance` in the g band, which helps to distinguish them from E/CEP/DSCT, while c-type RRL have values of `Meanvariance` in both bands similar to those of E/CEP/DSCT, which makes it difficult to tell them apart. The RRL class in our labeled set is dominated by ab-type RRL ($\sim$80%), and thus the light curve classifier identifies those more easily. However, for sources with $g \leq 15$, the fraction of ab-type RRL decreases to 64%, which explains the low recall values obtained for this regime of brightness.

In spite of all this, the fraction of RRL with only g band photometry is low (see Figure 5), and thus, the low performance obtained when we only consider the g band features for bright RRL does not substantially affect the presented results. As a comparison, when both bands are available for bright RRL, we classify %87.5 of them as RRL, but when we hide the r band features, leaving only the g band features, we recover only %17.2 of the bright RRL.

Figure 22. Logarithm of the Multiband period versus the `Meanvariance 1` (g band, left) and the `Meanvariance 2` (r band, right), for bright RRLs (mean $g < 16$) split according to their sub-classes 'ab' (green) and 'c' (red), and for E, CEP, and DSCT (grey). We show a zoom in the area where most of the RRLs lie. The contours show the density of points of each RRL class.

C.3. *The particular case of CV/Nova*

In Section 5.2 we show that the recall curves of the CV/Nova class are close to zero when only the g band is available. We noticed that in this case most of the CV/Novae are misclassified as periodic by the top level of the classifier (i.e., the class with the highest probability is periodic), and that most of them are misclassified as CEP or RRL by the bottom level. For these sources, the second or third classes with the highest probabilities in the bottom level is CV/Nova, and the probability of the CV/Nova class returned by the Stochastic classifier $[P_S(CV/Nova)]$ is larger than the probability of being CEP or RRL returned by the Periodic classifier $[P_P(CEP)$ or $P_P(RRL)]$. This confusion of the CV/Nova class with the periodic classes can also be seen in Figures 7 and 12. Therefore, the problem is produced in the first level of the BRF model.

We inspected the feature distributions in both bands for the CV/Nova, CEP, and RRL classes, in order to understand why the results obtained when only the g band is available are so different from the results obtained when only the r band is available. We did not find large differences among the g and r feature distributions, except for the $g-$W3 and $r-$W3 colors, where the confusion of CV/Nova with CEP and RRL is larger for the case of $g-$W3, as can be seen in Figure 23. However, these differences are not large enough to totally explain the low recall curves obtained for the CV/Nova class. Another possible explanation is that there is a large fraction of Cepheids with photometry only in the

g band (see Figure 5). This may be playing a role in the obtained results, since for the BRF model a lack of features in the r band and AllWISE+ZTF colors similar to those of the periodic classes could imply that the source is periodic.

Despite all this, more than 96% of the CV/Novae have photometry available in both bands, thus the low recall obtained for the g band is not a relevant issue. We plan to explore in future work better ways to classify CV/Novae. In particular, we will focus our efforts on classifying sub-classes of CVs and Novae.

Figure 23. Normalized $g-$W3 (left) and $r-$W3 (right) distributions, for active galaxies (red), CEP (blue), RRL (cyan), and CV/Nova (yellow).

D. SKY DISTRIBUTION OF THE EXTRAGALACTIC CANDIDATES

Figure 24 shows the sky distribution (in Galactic coordinates) of extragalactic candidates (QSO, AGN, Blazar, SNIa, SNIbc, SNII, and SLSN). It is expected that only a few extragalactic candidates are observed in the Galactic plane, which is confirmed in Figure 24.

REFERENCES

Abadi, M., Barham, P., Chen, J., et al. 2016, in 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), 265–283

Abril, J., Schmidtobreick, L., Ederoclite, A. r., & López-Sanjuan, C. 2020, MNRAS, 492, L40, doi: 10.1093/mnrasl/slz181

Aguirre, C., Pichara, K., & Becker, I. 2019, MNRAS, 482, 5078, doi: 10.1093/mnras/sty2836

Allevato, V., Paolillo, M., Papadakis, I., & Pinto, C. 2013, ApJ, 771, 9, doi: 10.1088/0004-637X/771/1/9

Arévalo, P., Churazov, E., Zhuravleva, I., Hernández-Monteagudo, C., & Revnivtsev, M. 2012, MNRAS, 426, 1793, doi: 10.1111/j.1365-2966.2012.21789.x

Arnett, D. 2008, in American Institute of Physics Conference Series, Vol. 1053, American Institute of Physics Conference Series, ed. S. K. Chakrabarti & A. S. Majumdar, 237–242

Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33, doi: 10.1051/0004-6361/201322068

Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, AJ, 156, 123, doi: 10.3847/1538-3881/aabc4f

Baldeschi, A., Miller, A., Stroh, M., Margutti, R., & Coppejans, D. L. 2020, arXiv e-prints, arXiv:2005.00155. https://arxiv.org/abs/2005.00155

Becker, I., Pichara, K., Catelan, M., et al. 2020, MNRAS, 493, 2981, doi: 10.1093/mnras/staa350

Bellm, E. 2014, in The Third Hot-wiring the Transient Universe Workshop, ed. P. R. Wozniak, M. J. Graham, A. A. Mahabal, & R. Seaman, 27–33

Bellm, E. C., Kulkarni, S. R., Graham, M. J., et al. 2019, PASP, 131, 018002, doi: 10.1088/1538-3873/aaecbe

Ben-Israel, A., & Greville, T. 2003, Adi Ben-Israel and Thomas N.E. Greville, Generalized Inverses: Theory and Applications, ISBN 0-387-00293-6, doi: https://doi.org/10.1007/b97366

Figure 24. Galactic latitude (gal_b, in degrees) versus Galactic longitude (gal_l, in degrees) for extragalactic candidates (QSO, AGN, Blazar, SNIa, SNIbc, SNII, and SLSN classes). The contours show the density of points in the plot. The bottom level probability computed by the deployed BRF classifier are color-coded according to the color bar to the right.

Boone, K. 2019, AJ, 158, 257,
doi: 10.3847/1538-3881/ab5182

Breiman, L. 2001, Machine Learning, 45, 5,
doi: 10.1023/A:1010933404324

Butler, N. R., & Bloom, J. S. 2011, AJ, 141, 93,
doi: 10.1088/0004-6256/141/3/93

Calzetti, D., Armus, L., Bohlin, R. C., et al. 2000, ApJ,
533, 682, doi: 10.1086/308692

Caplar, N., Lilly, S. J., & Trakhtenbrot, B. 2017, ApJ, 834,
111, doi: 10.3847/1538-4357/834/2/111

Carrasco-Davis, R., Cabrera-Vives, G., Förster, F., et al.
2019, PASP, 131, 108006, doi: 10.1088/1538-3873/aaef12

Carrasco-Davis, R., Reyes, E., Valenzuela, C., et al. 2020,
arXiv e-prints, arXiv:2008.03309.
https://arxiv.org/abs/2008.03309

Castro, N., Protopapas, P., & Pichara, K. 2018, AJ, 155,
16, doi: 10.3847/1538-3881/aa9ab8

Catelan, M., Dekany, I., Hempel, M., & Minniti, D. 2013,
Boletin de la Asociacion Argentina de Astronomia La
Plata Argentina, 56, 153

Catelan, M., & Smith, H. A. 2015, Pulsating Stars
(Wiley-VCH, Weinheim)

Chambers, K. C., Magnier, E. A., Metcalfe, N., et al. 2016,
arXiv e-prints, arXiv:1612.05560.
https://arxiv.org/abs/1612.05560

Chen, C., Liaw, A., Breiman, L., et al. 2004, University of
California, Berkeley, 110, 24

Chen, T., & Guestrin, C. 2016, in Proceedings of the 22nd
ACM SIGKDD International Conference on Knowledge
Discovery and Data Mining, KDD '16 (New York, NY,
USA: ACM), 785–794.
http://doi.acm.org/10.1145/2939672.2939785

Chollet, F., et al. 2015, Keras, https://keras.io

De Cicco, D., Paolillo, M., Falocco, S., et al. 2019, A&A,
627, A33, doi: 10.1051/0004-6361/201935659

Debosscher, J., Sarro, L. M., Aerts, C., et al. 2007, A&A,
475, 1159, doi: 10.1051/0004-6361:20077638

Debosscher, J., Sarro, L. M., López, M., et al. 2009, A&A,
506, 519, doi: 10.1051/0004-6361/200911618

D'Isanto, A., Cavuoti, S., Brescia, M., et al. 2016, MNRAS,
457, 3119, doi: 10.1093/mnras/stw157

Drake, A. J., Djorgovski, S. G., Mahabal, A., et al. 2009,
ApJ, 696, 870, doi: 10.1088/0004-637X/696/1/870

Drake, A. J., Graham, M. J., Djorgovski, S. G., et al. 2014,
ApJS, 213, 9, doi: 10.1088/0067-0049/213/1/9

Drake, A. J., Djorgovski, S. G., Catelan, M., et al. 2017,
MNRAS, 469, 3688, doi: 10.1093/mnras/stx1085

Elorrieta, F., Eyheramendy, S., Jordán, A., et al. 2016,
A&A, 595, A82, doi: 10.1051/0004-6361/201628700

Eyheramendy, S., Elorrieta, F., & Palma, W. 2018,
MNRAS, 481, 4311, doi: 10.1093/mnras/sty2487

Flesch, E. W. 2015, PASA, 32, e010,
doi: 10.1017/pasa.2015.10

—. 2019, arXiv e-prints, arXiv:1912.05614.
https://arxiv.org/abs/1912.05614

Foley, R. J., & Mandel, K. 2013, ApJ, 778, 167,
doi: 10.1088/0004-637X/778/2/167

Förster, F., Maureira, J. C., San Martín, J., et al. 2016,
ApJ, 832, 155, doi: 10.3847/0004-637X/832/2/155

Förster, F., Moriya, T. J., Maureira, J. C., et al. 2018,
Nature Astronomy, 2, 808,
doi: 10.1038/s41550-018-0563-4

Förster, F., Cabrera-Vives, G., Castillo-Navarrete, E., et al.
2020, arXiv e-prints, arXiv:2008.03303.
https://arxiv.org/abs/2008.03303

Friedman, J. H. 2001, Annals of statistics, 1189

Graham, M. J., Djorgovski, S. G., Drake, A. J., et al. 2017,
MNRAS, 470, 4112, doi: 10.1093/mnras/stx1456

Graham, M. J., Drake, A. J., Djorgovski, S. G., et al. 2013,
MNRAS, 434, 3423, doi: 10.1093/mnras/stt1264

Guillochon, J., Parrent, J., Kelley, L. Z., & Margutti, R.
2017, ApJ, 835, 64, doi: 10.3847/1538-4357/835/1/64

Hastie, T., Tibshirani, R., & Friedman, J. 2009, The
Elements of Statistical Learning: Data Mining, Inference,
and Prediction, Springer series in statistics (Springer).
https:
//link.springer.com/book/10.1007/978-0-387-84858-7

Haykin, S. 1994, Neural networks: a comprehensive
foundation (Prentice Hall PTR)

Holoien, T. W. S., Brown, J. S., Vallely, P. J., et al. 2019,
MNRAS, 484, 1899, doi: 10.1093/mnras/stz073

Hosenie, Z., Lyon, R., Stappers, B., Mootoovaloo, A., &
McBride, V. 2020, MNRAS, 493, 6050,
doi: 10.1093/mnras/staa642

Hosenie, Z., Lyon, R. J., Stappers, B. W., & Mootoovaloo,
A. 2019, MNRAS, 488, 4858, doi: 10.1093/mnras/stz1999

Huijse, P., Estévez, P. A., Förster, F., et al. 2018, ApJS,
236, 12, doi: 10.3847/1538-4365/aab77c

Hunter, J. D. 2007, Computing in Science Engineering, 9,
90

Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873,
111, doi: 10.3847/1538-4357/ab042c

Jayasinghe, T., Kochanek, C. S., Stanek, K. Z., et al. 2018,
MNRAS, 477, 3145, doi: 10.1093/mnras/sty838

Jayasinghe, T., Stanek, K. Z., Kochanek, C. S., et al.
2019a, MNRAS, 486, 1907, doi: 10.1093/mnras/stz844

—. 2019b, MNRAS, 485, 961, doi: 10.1093/mnras/stz444

—. 2020, MNRAS, 491, 13, doi: 10.1093/mnras/stz2711

Jurić, M., Axelrod, T., Becker, A., et al. 2019, LSST
Document LSE-163, doi: https://ls.st/LSE-163

Kelly, B. C., Bechtold, J., & Siemiginowska, A. 2009, ApJ,
698, 895, doi: 10.1088/0004-637X/698/1/895

Kim, D.-W., & Bailer-Jones, C. A. L. 2016, A&A, 587,
A18, doi: 10.1051/0004-6361/201527188

Kim, D.-W., Protopapas, P., Bailer-Jones, C. A. L., et al.
2014, A&A, 566, A43, doi: 10.1051/0004-6361/201323252

Kim, D.-W., Protopapas, P., Byun, Y.-I., et al. 2011, ApJ,
735, 68, doi: 10.1088/0004-637X/735/2/68

Kluyver, T., Ragan-Kelley, B., Pérez, F., et al. 2016, in
Positioning and Power in Academic Publishing: Players,
Agents and Agendas, ed. F. Loizides & B. Schmidt, IOS
Press, 87 – 90

Komossa, S. 2015, Journal of High Energy Astrophysics, 7,
148, doi: 10.1016/j.jheap.2015.04.006

Kovács, G., Zucker, S., & Mazeh, T. 2002, A&A, 391, 369,
doi: 10.1051/0004-6361:20020802

Lemaître, G., Nogueira, F., & Aridas, C. K. 2017, Journal
of Machine Learning Research, 18, 1

MacLeod, C. L., Ivezić, Ž., Kochanek, C. S., et al. 2010,
ApJ, 721, 1014, doi: 10.1088/0004-637X/721/2/1014

Mainzer, A., Bauer, J., Grav, T., et al. 2011, ApJ, 731, 53,
doi: 10.1088/0004-637X/731/1/53

Martínez-Palomera, J., Förster, F., Protopapas, P., et al.
2018, AJ, 156, 186, doi: 10.3847/1538-3881/aadfd8

Masci, F. J., Laher, R. R., Rusholme, B., et al. 2019,
PASP, 131, 018003, doi: 10.1088/1538-3873/aae8ac

Massaro, E., Maselli, A., Leto, C., et al. 2015, Ap&SS, 357,
75, doi: 10.1007/s10509-015-2254-2

McKinney, W., et al. 2010, in Proceedings of the 9th
Python in Science Conference, Vol. 445, Austin, TX,
51–56

McLaughlin, M. A., Mattox, J. R., Cordes, J. M., &
Thompson, D. J. 1996, ApJ, 473, 763,
doi: 10.1086/178188

McWhirter, P. R., Steele, I. A., Hussain, A., Al-Jumeily,
D., & Vellasco, M. M. B. R. 2018, MNRAS, 479, 5196,
doi: 10.1093/mnras/sty1823

Metzger, B. D., Martínez-Pinedo, G., Darbha, S., et al.
2010, MNRAS, 406, 2650,
doi: 10.1111/j.1365-2966.2010.16864.x

Mighell, K. J., & Plavchan, P. 2013, AJ, 145, 148,
doi: 10.1088/0004-6256/145/6/148

Mondrik, N., Long, J. P., & Marshall, J. L. 2015, ApJL,
811, L34, doi: 10.1088/2041-8205/811/2/L34

Mowlavi, N., Lecoeur-Taïbi, I., Lebzelter, T., et al. 2018,
A&A, 618, A58, doi: 10.1051/0004-6361/201833366

Muthukrishna, D., Narayan, G., Mandel, K. S., Biswas, R.,
& Hložek, R. 2019, PASP, 131, 118002,
doi: 10.1088/1538-3873/ab1609

Narayan, G., Zaidi, T., Soraisam, M. D., et al. 2018, ApJS, 236, 9, doi: 10.3847/1538-4365/aab781

Naul, B., Bloom, J. S., Pérez, F., & van der Walt, S. 2018, Nature Astronomy, 2, 151, doi: 10.1038/s41550-017-0321-z

Netzer, H. 2013, The Physics and Evolution of Active Galactic Nuclei

Nun, I., Protopapas, P., Sim, B., & Chen, W. 2016, AJ, 152, 71, doi: 10.3847/0004-6256/152/3/71

Nun, I., Protopapas, P., Sim, B., et al. 2015, ArXiv e-prints. https://arxiv.org/abs/1506.00010

Oh, K., Yi, S. K., Schawinski, K., et al. 2015, ApJS, 219, 1, doi: 10.1088/0067-0049/219/1/1

Padovani, P., Alexander, D. M., Assef, R. J., et al. 2017, A&A Rv, 25, 2, doi: 10.1007/s00159-017-0102-9

Palanque-Delabrouille, N., Magneville, C., Yèche, C., et al. 2016, A&A, 587, A41, doi: 10.1051/0004-6361/201527392

Palaversa, L., Ivezić, Ž., Eyer, L., et al. 2013, AJ, 146, 101, doi: 10.1088/0004-6256/146/4/101

Paolillo, M., Schreier, E. J., Giacconi, R., Koekemoer, A. M., & Grogin, N. A. 2004, ApJ, 611, 93, doi: 10.1086/421967

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2012, arXiv e-prints, arXiv:1201.0490. https://arxiv.org/abs/1201.0490

Peters, C. M., Richards, G. T., Myers, A. D., et al. 2015, ApJ, 811, 95, doi: 10.1088/0004-637X/811/2/95

Rakshit, S., Stalin, C. S., & Kotilainen, J. 2020, ApJS, 249, 17, doi: 10.3847/1538-4365/ab99c5

Richards, G. T., Myers, A. D., Gray, A. G., et al. 2009, ApJS, 180, 67, doi: 10.1088/0067-0049/180/1/67

Richards, J. W., Starr, D. L., Miller, A. A., et al. 2012, ApJS, 203, 32, doi: 10.1088/0067-0049/203/2/32

Richards, J. W., Starr, D. L., Butler, N. R., et al. 2011, ApJ, 733, 10, doi: 10.1088/0004-637X/733/1/10

Rimoldini, L., Holl, B., Audard, M., et al. 2019, A&A, 625, A97, doi: 10.1051/0004-6361/201834616

Ritter, H., & Kolb, U. 2003, A&A, 404, 301, doi: 10.1051/0004-6361:20030330

Rokach, L., & Maimon, O. Z. 2008, Data mining with decision trees: theory and applications, Vol. 69 (World scientific)

Sánchez, P., Lira, P., Cartier, R., et al. 2017, ApJ, 849, 110, doi: 10.3847/1538-4357/aa9188

Sánchez-Sáez, P., Lira, P., Mejía-Restrepo, J., et al. 2018, ApJ, 864, 87, doi: 10.3847/1538-4357/aad7f9

Sánchez-Sáez, P., Lira, P., Cartier, R., et al. 2019, ApJS, 242, 10, doi: 10.3847/1538-4365/ab174f

Sánchez-Sáez, P., Reyes, I., Valenzuela, C., et al. 2020, doi: 10.5281/ZENODO.4279451. https://zenodo.org/record/4279451

Sarro, L., Debosscher, J., López, M., & Aerts, C. 2009, Astronomy & Astrophysics, 494, 739

Sartori, L. F., Trakhtenbrot, B., Schawinski, K., et al. 2019, ApJ, 883, 139, doi: 10.3847/1538-4357/ab3c55

Schmidt, K. B., Marshall, P. J., Rix, H.-W., et al. 2010, ApJ, 714, 1194, doi: 10.1088/0004-637X/714/2/1194

Schwarzenberg-Czerny, A. 1996, ApJL, 460, L107, doi: 10.1086/309985

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. 2014, Journal of Machine Learning Research, 15, 1929

Stellingwerf, R. F., & Donohoe, M. 1986, ApJ, 306, 183, doi: 10.1086/164331

Tachibana, Y., & Miller, A. A. 2018, PASP, 130, 128001, doi: 10.1088/1538-3873/aae3d9

Udalski, A., Szymanski, M., Kaluzny, J., Kubiak, M., & Mateo, M. 1992, AcA, 42, 253

van der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, Computing in Science Engineering, 13, 22

Van Rossum, G., & Drake Jr, F. L. 1995, Python reference manual (Centrum voor Wiskunde en Informatica Amsterdam)

VanderPlas, J. T. 2018, The Astrophysical Journal Supplement Series, 236, 16, doi: 10.3847/1538-4365/aab766

Villar, V. A., Berger, E., Miller, G., et al. 2019a, ApJ, 884, 83, doi: 10.3847/1538-4357/ab418c

—. 2019b, ApJ, 884, 83, doi: 10.3847/1538-4357/ab418c

Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, Nature Methods, 17, 261, doi: https://doi.org/10.1038/s41592-019-0686-2

Wang, C., Deng, C., & Wang, S. 2019, arXiv e-prints, arXiv:1908.01672. https://arxiv.org/abs/1908.01672

Waskom, M., Botvinnik, O., O'Kane, D., et al. 2017, mwaskom/seaborn: v0.8.1 (September 2017), v0.8.1, Zenodo, doi: 10.5281/zenodo.883859. https://doi.org/10.5281/zenodo.883859

Wenger, M., Ochsenbein, F., Egret, D., et al. 2000, A&AS, 143, 9, doi: 10.1051/aas:2000332

Woosley, S. E., Heger, A., & Weaver, T. A. 2002, Reviews of Modern Physics, 74, 1015, doi: 10.1103/RevModPhys.74.1015

Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868, doi: 10.1088/0004-6256/140/6/1868

Zorich, L., Pichara, K., & Protopapas, P. 2020, MNRAS, 492, 2897, doi: 10.1093/mnras/stz3426