Publication Year	2019
Acceptance in OA@INAF	2021-01-15T16:26:03Z
Title	Gaia Data Release 2. All-sky classification of high-amplitude pulsating stars
Authors	Rimoldini, L.; Holl, B.; Audard, M.; Mowlavi, N.; Nienartowicz, K.; et al.
DOI	10.1051/0004-6361/201834616
Handle	http://hdl.handle.net/20.500.12386/29792
Journal	ASTRONOMY & ASTROPHYSICS
Number	625

A&A 625, A97 (2019)
https://doi.org/10.1051/0004-6361/201834616
© ESO 2019

Astronomy & Astrophysics

Gaia Data Release 2

Special issue

Gaia Data Release 2

All-sky classification of high-amplitude pulsating stars

L. Rimoldini[1], B. Holl[1,2], M. Audard[1,2], N. Mowlavi[1,2], K. Nienartowicz[3], D. W. Evans[4], L. P. Guy[1,5], I. Lecoeur-Taïbi[1], G. Jevardat de Fombelle[3], O. Marchal[1,6], M. Roelens[1,2,7], J. De Ridder[8], L. M. Sarro[9], S. Regibo[8], M. Lopez[10], G. Clementini[11], V. Ripepi[12], R. Molinaro[12], A. Garofalo[11], L. Molnár[13,14], E. Plachy[13,14], Á. Juhász[13,15], L. Szabados[13], T. Lebzelter[16], D. Teyssier[17], and L. Eyer[2]

[1] Department of Astronomy, University of Geneva, Chemin d'Ecogia 16, 1290 Versoix, Switzerland
e-mail: Lorenzo.Rimoldini@unige.ch
[2] Department of Astronomy, University of Geneva, Chemin des Maillettes 51, 1290 Versoix, Switzerland
[3] SixSq, Rue du Bois-du-Lan 8, 1217 Meyrin, Switzerland
[4] Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK
[5] AURA/LSST, 950 N Cherry Avenue, Tucson, Arizona 85719, USA
[6] GÉPI, Observatoire de Paris, Université Paris Sciences & Lettres, CNRS, Place Jules Janssen 5, 92195 Meudon, France
[7] Institute of Global Health, 9 Chemin des Mines, 1202 Geneva, Switzerland
[8] Institute of Astronomy, KU Leuven, Celestijnenlaan 200D, 3001 Leuven, Belgium
[9] Departamento Inteligencia Artificial, Universidad Nacional de Educación a Distancia, Calle Juan del Rosal 16, 28040 Madrid, Spain
[10] Departamento de Astrofísica, Centro de Astrobiología (INTA-CSIC), PO Box 78, 28691 Villanueva de la Cañada, Spain
[11] INAF – Osservatorio di Astrofisica e Scienza dello Spazio di Bologna, Via Gobetti 93/3, 40129 Bologna, Italy
[12] INAF – Osservatorio Astronomico di Capodimonte, Via Moiariello 16, 80131 Napoli, Italy
[13] Konkoly Observatory, Research Centre for Astronomy & Earth Sciences, Hungarian Academy of Sciences, Konkoly Thege Miklós út 15–17, 1121 Budapest, Hungary
[14] MTA CSFK Lendület Near-Field Cosmology Research Group, Konkoly Thege Miklós út 15–17, 1121 Budapest, Hungary
[15] Department of Astronomy, Eötvös Loránd University, Pázmány Péter sétány 1/a, 1117 Budapest, Hungary
[16] Department of Astrophysics, University of Vienna, Tuerkenschanzstrasse 17, 1180 Vienna, Austria
[17] Telespazio Vega UK Ltd for ESA/ESAC, Camino bajo del Castillo, s/n, Urbanizacion Villafranca del Castillo, Villanueva de la Cañada, 28692 Madrid, Spain

Received 9 November 2018 / Accepted 17 March 2019

ABSTRACT

Context. More than half a million of the 1.69 billion sources in *Gaia* Data Release 2 (DR2) are published with photometric time series that exhibit light variations during the 22 months of observation.

Aims. An all-sky classification of common high-amplitude pulsators (Cepheids, long-period variables, δ Scuti/SX Phoenicis, and RR Lyrae stars) is provided for stars with brightness variations greater than 0.1 mag in G band.

Methods. A semi-supervised classification approach was employed, firstly training multi-stage random forest classifiers with sources of known types in the literature, followed by a preliminary classification of the *Gaia* data and a second training phase that included a selection of the first classification results to improve the representation of some classes, before the improved classifiers were applied to the *Gaia* data. Dedicated validation classifiers were used to reduce the level of contamination in the published results. A relevant fraction of objects were not yet sufficiently sampled for reliable Fourier series decomposition, consequently classifiers were based on features derived from statistics of photometric time series in the G, $G_{\rm BP}$, and $G_{\rm RP}$ bands, as well as from some astrometric parameters.

Results. The published classification results include 195 780 RR Lyrae stars, 150 757 long-period variables, 8550 Cepheids, and 8882 δ Scuti/SX Phoenicis stars. All of these results represent candidates whose completeness and contamination are described as a function of variability type and classification reliability. Results are expressed in terms of class labels and classification scores, which are available in the `vari_classifier_result` table of the *Gaia* archive.

Key words. catalogs – methods: data analysis – stars: variables: general – stars: variables: Cepheids – stars: variables: delta Scuti – stars: variables: RR Lyrae

1. Introduction

The light curves of variable stars exhibit features that can reveal valuable information on the physical causes of brightness variations. They help us improve our understanding of stellar properties, some of which can be used to develop and refine methods that turn variable stars into astrophysical tools applicable to Galactic or even extra-galactic scales (Aerts & Sterken 2006).

The brightness variations of pulsating stars are caused by periodic expansion and contraction, which may alternate throughout the whole star (causing a uniform swelling, shrinking, and temperature changes across the entire stellar surface)

or in the case of non-radial oscillations occur simultaneously but in different regions of the star (typically associated with lower brightness variations). The amplitude, periodicity, regularity, lifetime, and other features of stellar pulsations depend on stellar evolution stages, which can be mapped on the Hertzsprung–Russell diagram and are found to correspond to different observed types of variability (e.g. Gaia Collaboration 2019), which are commonly identified by the light-curve shapes, pulsation period(s), amplitude(s), and intrinsic colours, among other factors.

Pulsating variables became particularly important for the information that can be inferred on stellar interiors with asteroseismology (e.g. Aerts et al. 2010; Christensen-Dalsgaard 2004) and the relationships between the pulsation periods and stellar luminosities (e.g. Sandage & Tammann 2006; Gaia Collaboration 2017; Clementini et al. 2019), which can be used to determine the distance, and the three-dimensional distributions of these stars can thus outline the structures they belong to within our Galaxy and beyond (e.g. Hernitschek et al. 2017; Drake et al. 2013a).

Examples of pulsating variables in recent surveys include the Panoramic Survey Telescope and Rapid Response System (Pan-STARRS; Sesar et al. 2017; Hernitschek et al. 2016), Catalina (Torrealba et al. 2015; Drake et al. 2014, 2013a,b), the Optical Gravitational Lensing Experiment (OGLE; Soszyński et al. 2017, 2016a, 2015a,b, 2014), *Kepler* (Bradley et al. 2015; Debosscher et al. 2011), the Asteroid Terrestrial-impact Last Alert System (ATLAS; Heinze et al. 2018), *Gaia* DR1 (Eyer et al. 2017; Clementini et al. 2016), the Lincoln Near-Earth Asteroid Research (LINEAR; Palaversa et al. 2013), the Northern Sky Variability Survey (NSVS; ROTSE Collaboration 2006; Woźniak et al. 2004), the All Sky Automated Survey (ASAS; Pojmanski 2002), and ASAS for Supernovae (ASAS-SN; Jayasinghe et al. 2018, 2019). Surveys that cover large regions of the sky tend to collect remarkable volumes of data, which help improve our understanding of known objects, increase the knowledge of rare objects, enable discoveries, and raise new questions, while providing all the data in a consistent context. *Gaia* pursues all this across the whole sky with the benefits of a rich set of instruments (Gaia Collaboration 2016) that repeatedly provide the astrometry, photometry, and spectroscopy of the observed objects. This increases the measurement accuracy and follows variations in time.

The second Data Release (DR2) of *Gaia* (Gaia Collaboration 2018) includes an earlier-than-planned publication of variable sources and related photometric time series in the G, $G_{\rm BP}$, and $G_{\rm RP}$ bands. In particular, 550 737 variable stars are published (as summarised in Holl et al. 2018), among which 363 969 objects are identified by the all-sky classification pipeline (see Sect. 7.3 of Eyer et al. 2018) as candidate RR Lyrae stars, Cepheids, δ Scuti/SX Phoenicis stars, and long-period variables, which constitute the first published classification results from the *Gaia* variability pipeline, covering the whole sky and without a priori selections of sources based on their sampling (unless only a single measurement was available). Other variable stars published in *Gaia* DR2 originated from special variability detection algorithms (Lanzafame et al. 2018; Roelens et al. 2018) and from another classification of sources dedicated to well-sampled sources (see Sect. 7.2.3.6 in Eyer et al. 2018). The list of *Gaia* DR2 articles related to the data processing and analysis or validation of the published variable objects is presented in Holl et al. (2018).

The *Gaia* variability processing and analysis includes several modules dedicated to different tasks, as described in Eyer et al. (2017), such as the computation of statistical parameters, the detection of variability, the characterisation of light curves, the classification of variable objects, and specific object studies (SOS) that confirm the identifications from previous stages (and in some cases reclassify) and refine the description of variability of specific types. The large number of sources and the sparse sampling of *Gaia* time series do not allow for "perfect" classifications of all sources. Classification results of types that are verified by SOS are published for transparency of the pipeline processing, but also to make it possible to publish results of sources that SOS could not confirm or that were improperly rejected for various reasons (e.g. because of spurious information, insufficient observations or photometric precision, inaccurate astrometric parameters, or other factors). Consequently, classification results are generally more complete but also more contaminated and associated with a less accurate light-curve characterisation than the one derived from SOS.

For *Gaia* DR2, two independent classifications of variable objects were performed: one with at least 20 field-of-view (FoV) transits in the G band (described in Sect. 7.2.3.6 in Eyer et al. 2018), which included period search and Fourier modelling in the computation of classification attributes, and one that required at least 2 FoV transits in the G band, which covered the whole sky more homogeneously but was limited to statistical parameters to characterise source features. The latter was published in the *Gaia* DR2 archive and is described in detail in this article; a general overview and some of its technical details are presented also in Sect. 7.3 of Eyer et al. (2018). The (unpublished) classification that made use of modelling attributes added a total of 37 016 variable stars that were confirmed by different SOS packages. Thus, for *Gaia* DR2, the published SOS results were not strictly a subset of the published classifier results (see Fig. 3 in Holl et al. 2018, for an illustration per variability type).

The quality of the classification results was limited by time constraints and by the dependence on data produced by *Gaia* pipelines that were executed in parallel to ours, which in some cases forced us to use preliminary (not published) *Gaia* data products and reduced the options available for variability processing. Without restricting the publication of classifications to objects associated with well-sampled and fully understood time series (according to current knowledge), the effect of selection biases is reduced and the community is granted early access to additional time series, adding discovery potential and facilitating progress in currently debated topics (e.g. Belokurov et al. 2017; Jacyszyn-Dobrzeniecka & the OGLE Team 2018). In addition to the classification scores and other indications given in this article, subsets of sources with enhanced reliability and sufficient sampling are selected by subsequent pipeline modules (SOS) for some of the variability types, whose detailed analyses and validations are presented in separate articles (Mowlavi et al. 2018, for long-period variables, and Clementini et al. 2019 and Molnár et al. 2018, for Cepheids and RR Lyrae stars).

This article emphasises the method employed for the classification of stars covering the whole sky and presents its results with indications relevant to their usage. Following the internal agreement of the *Gaia* data processing and analysis consortium, no scientific exploitation of these results is made herein. The *Gaia* data set is summarised in Sect. 2, the method employed to identify four main classes of pulsating variables is described in Sect. 3, results and comparisons with the literature are outlined in Sect. 4, and conclusions are drawn in Sect. 5. Appendix A recalls the classes of the objects employed in the training set,

and Appendix B provides a few sample queries applicable to classification-related searches in the *Gaia* archive[1].

2. Data

The data of *Gaia* DR2 are accessible from the ESA *Gaia* archive, from four partner data centres[2], and from a number of affiliated data centres around the world. Due to time constraints and to multiple *Gaia* modules preparing for DR2 in parallel, only some of the published data products were available at the time of variability processing. The processing of variable sources employed photometric data of FoV transits (Evans et al. 2018; Riello et al. 2018) in the G, $G_{\rm BP}$, and $G_{\rm RP}$ bands, with a small fraction of per-CCD G-band photometry (Roelens et al. 2018), and a preliminary version of the astrometric solution (the final version is described in Lindegren et al. 2018), while it could not make use of (preliminary nor published) spectroscopic data (Sartoretti et al. 2018), radial velocities (Katz et al. 2019), and astrophysical parameters (Andrae et al. 2018). Moreover, the initial stages of the all-sky classification activities (the cross match of *Gaia* data with catalogues of variable objects from the literature, described in Sect. 3.1.1, and the determination of parameters for quasar attributes, mentioned in Sect. 3.1.3) relied on preliminary per-FoV photometry in the G, $G_{\rm BP}$, and $G_{\rm RP}$ bands in order to gain time for subsequent (more critical) classification phases. Despite these short-comings, we believe that the advantages of an early publication of time series outweighs the benefits of a fully consistent data product, as the published classifications are not critically affected and the community can start working on these time series (together with the other DR2 data products) without waiting for the next data release. Eventually, these inconsistencies and their effects on results are expected to be reduced (or vanish) in the future.

The time series were reconstructed from FoV transit data with times referred to the solar system barycentre (Barycentric Coordinate Time), which means that they do not encode the motion of *Gaia* around the Sun (nor the time dilation from the gravitational potential of the latter). All sources with at least two FoV transits in the G band were processed, and suspicious time-series measurements were filtered out as outlined in Holl et al. (2018) and detailed in Sect. 7.2.3.2 of Eyer et al. (2018). The final *Gaia* validation stage, subsequent to the variability pipeline processing, removed duplicated sources and objects affected by imprecise astrometric solutions (Arenou et al. 2018), which increased the minimum number of G-band FoV transits of the published *Gaia* variables to five. The global properties of the published variable sources, an overview of their statistical parameters, magnitude and sky distributions per variability type, and the sky coverage as a function of the number of FoV transits per source are shown in Holl et al. (2018).

Section 3 describes the methods and their implementation for the classification of the *Gaia* variable objects across the whole sky, while Sect. 4 describes these classification results as they appear in the *Gaia* archive (i.e. after the validation cuts that followed the variability processing).

3. Methods

In the current era of big data, it is not possible to inspect every single light curve (at least not within the human resources and time allocated for the *Gaia* data releases). Machine-learning gives the possibility of automating decisions and thus processing a large number of sources based on patterns that are automatically recognised from much smaller controlled samples of training objects.

The initial goal for an advance classification across the whole sky with the *Gaia* DR2 data (thus including poorly sampled sources) was represented by RR Lyrae stars because they are important as standard candles and stellar population tracers (e.g. Clementini 2014, and references therein), and because they are easily identified through their large light variation amplitudes and short periods (typically shorter than a day), as other works classifying RR Lyrae stars with a low number of observations have shown (e.g. Ivezić et al. 2000; Sesar et al. 2007). The classifier developed for RR Lyrae stars included several variability types by design (to reduce the contamination of the targeted class). After an assessment of its results, it was decided that it was worth to extend the publication of the classification results to additional types of pulsating stars (Cepheids, δ Scuti/SX Phoenicis stars, and long-period variables) so that the community could further benefit from the publication of their time series as well. As this decision was reached after the construction of the classifier model, the training sources of these additional variability types were not as representative in the sky and in the magnitude distributions as the RR Lyrae training sample, and thus the associated classification results were more likely to be affected by training biases.

As described in Eyer et al. (2017), the classification of variable stars in the *Gaia* pipeline was based on attributes that characterised the objects (to train and then classify), such as the statistics of photometric time series, the associated modelling parameters, the astrometric properties, and other features if available. The classification results published in *Gaia* DR2 were obtained by means of supervised classification, which depended crucially on the selection of training-set objects (whose rationale is explained in Sect. 3.1), on the combination of attributes that best described the distinctive features of the related classes, although without period information (Sect. 3.1.3), and on the organisation of the classifier (Sect. 3.2). The implementation details of these procedures applied to various classifiers are presented in Sect. 3.2.2, followed by an outline of the automated validation of the first classification results (Sect. 3.3) and the definition of the classification score (Sect. 3.4). A summary of the steps described in this section is shown in Fig. 1, which presents the general flow of various procedures, while the details and exceptions are explained in the text. The classification quality estimators often mentioned in this work include the per-class completeness (i.e. true positive or recall) rates and contamination (i.e. 1 − precision or 1 − purity).

3.1. Training set

In supervised classification, the importance of training-set objects of known classes is paramount because results will be as good, poor, or biased as the training set. The data to train a classifier should be as similar as possible to the unlabelled data that are to be classified. In order to embed the *Gaia* properties (time sampling, photometric bands, and some of the data imperfections) in the training set, the latter was built using *Gaia* sources of independently known classes that resulted from the cross match of known objects in the literature with *Gaia* sources. The cross match with a large number of surveys, catalogues, sources, and variability types from the literature was necessary to gather sufficient objects from which training sources could be selected, and included a multitude of classes, covered all relevant sky

[1] http://gea.esac.esa.int/archive
[2] https://www.cosmos.esa.int/web/gaia/data-access

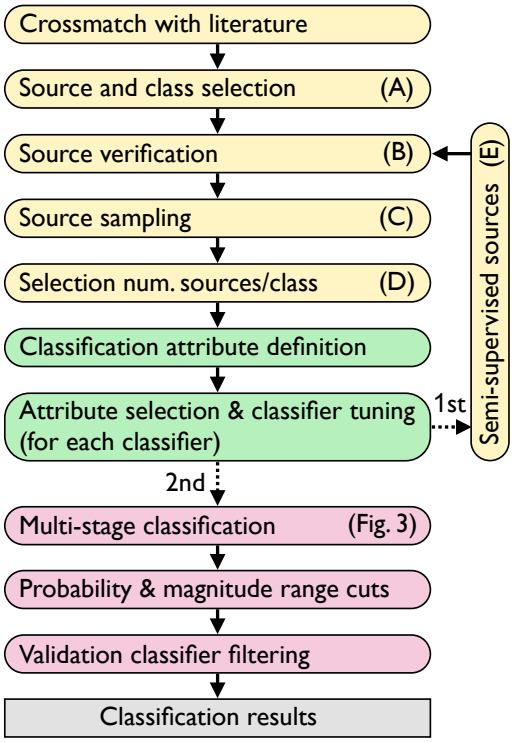


Fig. 1. General flow of the procedures employed to achieve the classification results. Labels (A)–(E) refer to the items described in Sect. 3.1.2. Exceptions (as in the case of the long-period variables, which were not filtered by a validation classifier) are described in the text. The box colours group the procedures by topic: training source/class selection (yellow), classification attributes (green), and classifiers and result selection (red).

regions per class (to probe location-dependent effects), and have magnitudes, colours, brightness variation amplitudes, periods (if periodic), number of FoV transits, and other parameters that are distributed according to the full range admitted for each class, convolved with the detectability of *Gaia*. It is usually not possible to achieve this level of detail for every parameter and class, but it remains important to pursue it within the available resources in order to limit the effect of training-set biases.

The representation of classes as a function of the number of FoV transits or sampling could be improved for those classes that suffered from insufficient sky coverage (which also reduced the diversity of time-series sampling), for example, by adding new training sources derived from downsampling better sampled time series. This approach was not employed in this work in order to preserve the attribute distributions and relations within the data, and also to prevent new biases and artefacts from the introduction of artificial sources. However, future classifications of *Gaia* variables might include such training extensions, depending on the outcome of analyses evaluating the benefits and costs of similar procedures for each class (e.g. Long et al. 2012).

3.1.1. Cross match

The classifier method that was employed to cross match variable objects with *Gaia* DR1 sources (Rimoldini et al. 2017) required significant preparatory work for each data set from the literature. Considering the large number of catalogues of variable sources targeted for cross match in the variability processing of *Gaia* DR2, a simplified method was developed to keep using multiple dimensions in order to reduce the number of incorrect and missed matches, but without the aid of machine-learning. A single multi-dimensional distance was built from different metrics, depending on their availability: angular separation (for all cases), colour (represented by the difference of time-series medians in the $G_{\rm BP}$ and $G_{\rm RP}$ bands for the *Gaia* data), magnitude (computed as the median time-series magnitude in the G band for the *Gaia* data), and amplitude estimators such as the range or standard deviation of G-band time-series magnitudes. These photometric parameters were computed after a simple filtering of *Gaia* FoV transits to reduce the possible influence of outliers: magnitudes above or below 15 times the median absolute deviation (MAD) from the median were removed (for all bands) and the top and bottom 10% of magnitudes in the G-band time series were filtered out for the amplitude estimation.

In order to compute a multi-dimensional distance, the metrics derived from *Gaia* photometric time series were compared with those from data sets of other surveys (with different bands) by means of empirical relations, limited to linear models and established iteratively: in the first iteration, matches were derived simply from the nearest neighbours (only astrometry) within one arcsecond, making the first cross-survey empirical relations available for use and refinement in subsequent iterations (that were not limited to astrometric comparisons, nor to the same cross match radius). For the second and subsequent iterations, all neighbours within a preset radius (typically a few arcseconds, depending on the survey) were considered as potential matches. A single match for each source was selected according to the smallest multi-dimensional distance[3], after the removal of likely duplicate *Gaia* sources (Arenou et al. 2018). If the same match was associated with multiple sources (as may happen in crowded fields), it was assigned to the source with the smallest cross-match distance, and new potential matches (if available) were (re-)considered for the other sources (and the same procedure was applied recursively in case of multiple sources for each match). The distributions of the differences of all pairs of metrics of the final iteration were visually inspected for the possibility of applying further constraints and excluding suspicious outliers. The time series of matched *Gaia* sources were plotted for verification, adjustments of the thresholds of the metrics, and their possible reassessment.

This method enabled us to cross match efficiently about 70 different data sets from the literature (for a total of about two million sources of known classes). The subset of catalogues considered sufficiently reliable for training purposes included over 750 thousand objects from the data sets listed in Table 7.1 of Eyer et al. (2018), together with literature references and the cross-match metrics employed. Although this large number of cross-matched sources included 29 classes, the constant stars, eclipsing binaries, and RR Lyrae variables accounted for 92% (about 54, 19, and 19%, respectively) of all sources; the largest contributions came from the Sloan Digital Sky Survey standard star catalogue for Stripe 82 (Ivezić et al. 2007), from the OGLE-IV eclipsing binaries and RR Lyrae stars identified in the Galactic bulge (Soszyński et al. 2016b, 2014) and in the Magellanic Clouds (Pawlak et al. 2016; Soszyński et al. 2016a), and from the Pan-STARRS1 classification of RR Lyrae stars (Sesar et al. 2017). These catalogues accounted for about 84% of all cross-matched objects. In the special case of OGLE-IV GSEP constant objects (Soszyński et al. 2012), no specific catalogue for

[3] The cross-match distance was defined as the sum in quadrature of the differences of metric values of two surveys; each component was normalised by the MAD of the distribution of the differences of the related metric that was derived from the matches of the previous iteration.

constant stars is available. About 10 thousand constant star candidates were derived from sources with the smallest variations in both V and I bands as a function of magnitude, after objects were removed because they were considered less reliable (with $V-I$ bluer than -1 mag or redder than 3 mag, with fewer than 20 or 200 observations in the V and I band, respectively, with measurements flagged as "bad", with standard deviation greater than 0.1 mag in either band, and fainter than 22 or 21 mag in the V and I band, respectively) or they were identified as galaxies by OGLE.

3.1.2. Classes and sources

The selections of classes and of training sources that represent them are pivotal to the purpose of the training set. In supervised classification, it is important to train with a variety of classes to reduce the chance of contamination from sources that belong to classes that are missing in the training set. In addition to the classes targeted for publication, those that are commonly confused with them as well as others that are well separated are expected to be part of a training set. Objects of untrained classes will be classified as one of the trained classes, possibly associated with low classification scores and/or significant anomaly or outlier indicators (such as the proximity[4] in random forest or the Mahalanobis distance, Mahalanobis 1936, in Gaussian mixture classifiers), which can help in the rejection of false positives. As a consequence of the literature data sets that were selected to classify variable stars for *Gaia* DR2, different combinations of 29 classes (see Sect. 3.2.1) were included in the training set with at least 10 representatives per class for a minimum class definition (although poorly sampled classes remained poorly defined and likely biased). These classes comprised constant and variable objects, whose labels and definitions were listed in Sect. 7.3.3.1 of Eyer et al. (2018); they are replicated herein for convenience in Appendix A. They include 11 subclasses that belonged to four families of pulsating stars (Cepheids, long-period variables, RR Lyrae, and δ Scuti/SX Phoenicis stars), which were published in the classification results of variables in the ESA *Gaia* archive[1].

The large number of cross-matched sources allowed us to be very selective and better control the distribution of objects in the sky, in magnitude, and in the number of G-band FoV transits for several classes and consequently reduce the effect of training-set biases on classifications. The final training set included about 33 thousand *Gaia* sources. Their selection involved the following semi-automated procedures interleaved with visual inspections to assess the quality of each stage and the subsequent steps.

(A) Literature data. After uniforming the class labels of different cross-matched data sets, the latter were first selected and then prioritised according to evaluations of their classification reliability, in order to resolve contradictions of different literature classes associated with the same *Gaia* source (as found in 0.2% of all matches, which included types that described different manifestations of the same phenomenon, such as flaring and rotating spotted stars). The reliability of literature classifications was based on survey features, on the location of sources in the sky, on the appearance of their light curves in the *Gaia* photometry (limited to samples of sources with contradictory classes in the literature), and on qualitative rankings from the experience of a subset of co-authors in using these data sets for some classes. This assessment did not correct for every classification disagreement in the literature, but it was deemed to reduce the effect of the remaining incorrect class assignments well below the effect of other sources of classification confusion.

(B) Source verification. Simple statistics (such as the median colour and magnitude, the range of magnitude variations, the skewness, and the Abbe value computed on magnitudes sorted in time and in phase with the literature period when available) were computed from the *Gaia* time series, and sources that did not satisfy constraints typical of their own class were rejected. These constraints were typically applied to two-dimensional projections of the statistics mentioned above for each trained class, but they remained rather permissive (i) to allow for a wide range of possible distance, extinction, and reddening (with consequent effects on magnitudes and colours), (ii) to consider the limited sampling of a good fraction of *Gaia* sources (with few FoV transits), and (iii) to include the occasional effect of spurious measurements that remained after the initial time-series cleaning (so that the classifier could expect such artefacts as well). In the case of eclipsing binaries, the reduced and sparse sampling might miss the signal, therefore a positive skewness was required in addition to a minimum G-band range of 0.1 mag.

(C) Sampling. Cross-matched sources incorporate the footprints of the original data sets, which can cause artificial (survey-specific) peaks in the distribution of several parameters. In order to alleviate predictable training biases, subsets of cross-matched sources were selected for more representative distributions in the sky, in magnitude, and in the number of FoV transits (contemporaneously), pursued separately for each class, and conditioned on the availability of at least about 10 000 sources per class (which was possible for classes denoted by the following labels, sorted by increasing occurrence: CEP, ELL, ROT, RRC, RRAB, ECL, and CONSTANT).

(D) Classifier priors. The classification method described in Sect. 3.2 was based on a set of random forest classifiers (Breiman 2001) that were generally sensitive to the number of sources for each class relative to each other. For better control of the relative importances of classes, the sources of each class were further selected up to preset quotas for each class. This selection took into account the literature origins of sources and applied different caps to the random draw of sources from specific cross-matched data sets in order to favour the representation of a variety of information in the training set and reduce the influence of large surveys whose properties or targeted locations only partially matched those of the *Gaia* data.

(E) Semi-supervised sources. Despite the large number of cross-matched sources, their number density (per class) decreased quickly in multi-dimensional volumes, leaving under-represented intervals in several cases. For the classes labelled CONSTANT, MIRA, RRAB, RRC, and SR, gaps in the magnitude and/or sky distributions could be filled with sources classified as such classes from a first execution of the classification module. The latter made no use of the attributes whose distributions were meant to be corrected (such as the G magnitude or sky location), nor of the attributes that were obviously correlated with them (such as the $G_{\rm BP}$ and $G_{\rm RP}$ magnitudes, inter-band correlations, and all astrometric attributes) in order to reduce classifier biases, although other attributes might have suffered from biases as well (some of which may be related to residual correlations with the targeted attributes as well). The

[4] https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

remaining attributes were selected according to the techniques described in Sect. 3.1.3.

In this semi-supervised approach, classified sources within magnitude intervals and sky regions that lacked representation were added to the training set if their classification confidence and the associated class completeness rate were estimated sufficiently high (to limit the chance of introducing contaminants). Such a dependence on the classification reliability might bias the sampling of sources. To mitigate this (less severe) effect and make the contribution of new training information possible, the classification confidence of new sources was also constrained by an upper limit. The same values of uncalibrated classification "probabilities" had different meaning for different classes, therefore these thresholds were assessed as a function of class. The new candidate training sources were verified by the same per-class statistics that were applied to the other training sources (item B), sampled to smooth distributions (item C, when applicable), or to fit the classifier priors (item D), and finally amounted to 11% of the training set globally (in particular, 14, 50, 13, 9, and 71% of CONSTANT, MIRA, RRAB, RRC, and SR classes, respectively). Reliable *Gaia* classifications are expected to increase their footprint (in terms of number of classes and contribution to each of them) in the training sets of variable objects in future *Gaia* data releases.

Figure 2 illustrates an example of the distributions of stars in the sky and in the *G*-band magnitude as they appeared in the cross match and in the training set in the particular case of the fundamental-mode RR Lyrae (RRAB) stars. The reduction of the number of sources and the smoothing of density peaks in both sky and magnitude distributions, based on the multi-dimensional source sampling described in item (C), are clearly visible in Fig. 2. Moreover, the addition of semi-supervised sources, following the procedure detailed in item (E), is shown to fill a region of the sky that previously lacked known representatives.

The final composition of the training-set classes included super- and sub-classes according to the hierarchical classifier organisation presented in Fig. 3, and their representation was provided together with the assessments of classifier models in Figs. 4–6 (because the number of sources for each class and the competing classes are meaningful in the context of each classifier and the classes targeted for publication).

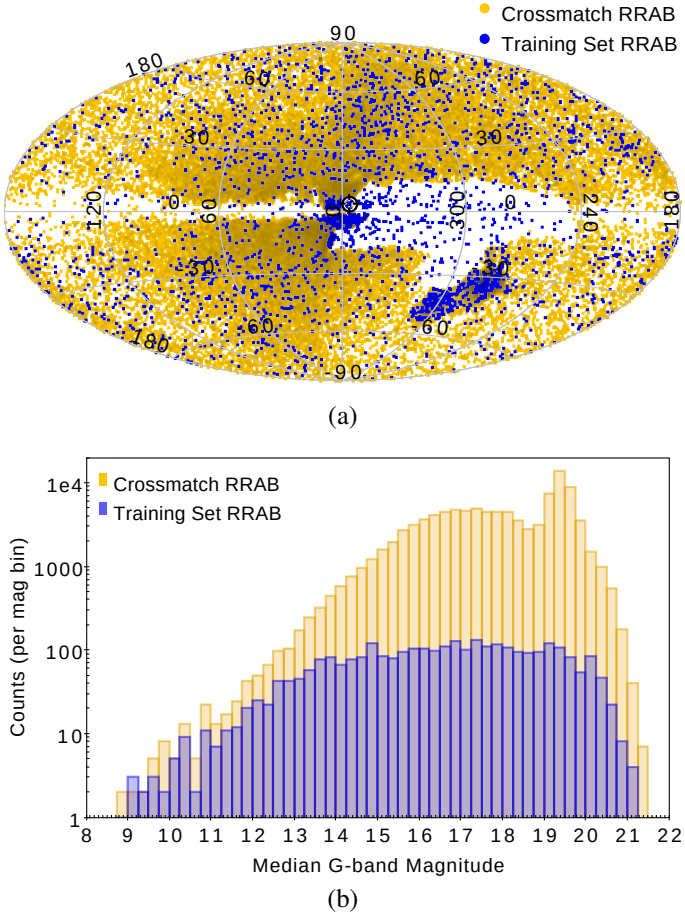


Fig. 2. Distributions of cross-match (orange) vs. training-set (blue) sources of RRAB type in the sky (*panel a*), in Galactic coordinates (degrees), and in the *G*-band magnitude (*panel b*). The source sampling helped reduce the number of sources for training and smoothed both magnitude and sky distributions (with intended over-densities in the regions of the Galactic bulge and Magellanic Clouds), while semi-supervised sources filled the under-represented region in the cross match centred around the Galactic longitude of 300°.

3.1.3. Attributes

Classification attributes describe the source features that help us distinguish the classes to which the sources belong. The extraction of such attributes may involve simple or complicated methods, statistics, models, and data types of different nature (e.g. photometric, astrometric, spectroscopic, global, or epoch specific). The translation of all this information into numerical values in a homogeneous way and in a common context makes it possible to compare sources, identify attributes with typical values for specific classes and other attributes (or combinations thereof) with values that differ the most for sources of different classes. Class models are defined by algorithms that are based on the classification attributes of training sources and are subsequently used to classify unlabelled objects using the same set of attribute definitions.

The effectiveness of attributes is determined by the relevance of the features to describe, convolved with data properties (such as sampling, accuracy, and rate of spurious measurements), which may also depend on source characteristics such as sky location, brightness, and others, making it difficult to guess a priori the most efficient attribute definitions. In the attempt to capture all possible features of each class and at the same time include different expressions for the same features, the number of attributes can grow quickly. Attributes are necessary elements of classification models, but too many of them can have an adverse effect. Although data sample sizes have increased enormously in recent years, the volume of the attribute space grows rapidly as the dimensionality increases, and the number density of sources represented in such volumes (and their statistical significance) decreases just as rapidly (this effect is also known as the curse of dimensionality, see Bellman 1961; Hastie et al. 2009). Too many attributes can also lead to overfitting the training-set features, so that the model loses generality when applied to unlabelled data. Even though the susceptibility to overfitting depends on the classification technique too (and random forest proved to be one of the most robust methods in this context), a smaller number of attributes can improve the model in terms of learning accuracy, general applicability, and interpretability, in addition to lower computational requirements (i.e. time and storage resources).

Common techniques to reduce dimensionality include the combination of multiple attributes and/or the selection of a subset of highly discriminant attributes that optimally split different classes. We focussed on the latter and tested different

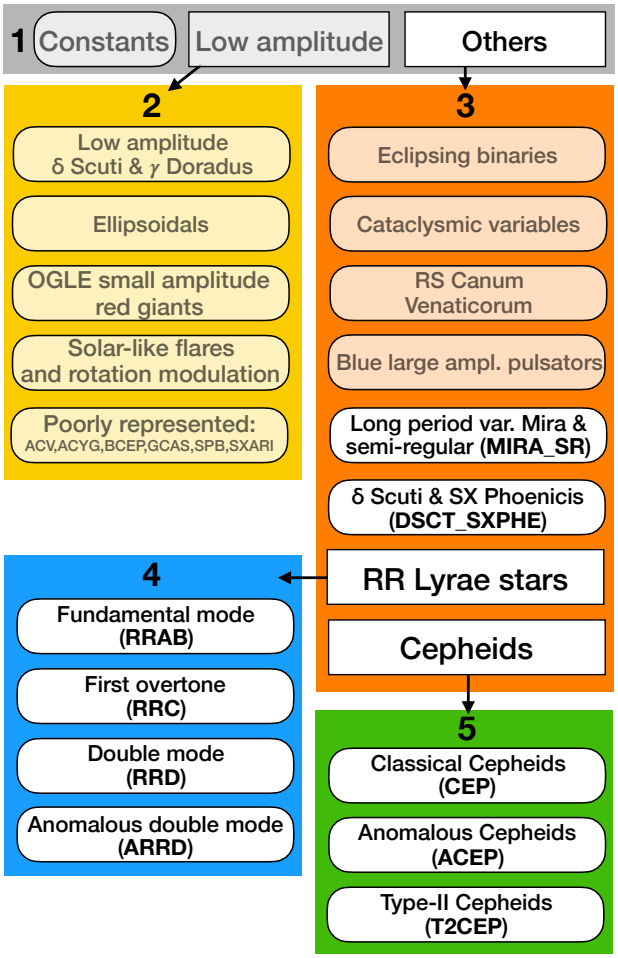


Fig. 3. Multi-stage classification tree of five random forest classifiers. Each classifier is identified by a number within the shaded region of the same colour that encompasses the (sub)types of objects to classify. The square boxes denote superclasses that are further split in subsequent stages. The names of the final classification types (which include subtypes and class combinations) appear in round-corner boxes, and only those with a white background are published in *Gaia* DR2.

methods to efficiently optimise the training sets of five classifiers (described in Sect. 3.2) that initially contained 150 attributes for a total of 33 thousand training sources. Some of these techniques were implemented in the literature (Calle et al. 2011; Genuer et al. 2010; Diaz-Uriarte & Alvarez de Andres 2005), while others were coded according to the principles of forward selection and backward elimination of attributes (Guyon & Elisseeff 2003), that is, by progressively adding the most useful attributes or removing the least useful ones (where the usefulness of each tested attribute was evaluated by the change in the classifier accuracy rate). The backward elimination can capture more synergies among attributes than the forward-selection technique (which evaluates only subsets of possible attribute combinations), but the former can be very demanding in terms of processing time, and computationally less intensive solutions are often pursued. Attributes are then ranked by their selection order, which rewards the truly useful ones and penalises the inefficient, noisy, and redundant ones. Every attribute selection method employed random forest as classification technique, and the trained model assessment was based on "out-of-bag" sources (unused training objects from the random draw with replacement to build each tree).

Given the time requirements of the iterative attribute testing on a relatively large volume of training data, the forward selection of the most useful attributes was limited to the identification of the top 12 attributes, while the backward elimination evaluated the classifier accuracy after each removal of the least important attributes and reassessed the importance of the remaining attributes after each iteration. The attribute importance is available for all attributes of a given random forest classifier (computationally less intensive than assessing the effect of each attribute by retraining classifiers without each of them) and is evaluated by the mean decrease in accuracy after shuffling the values of each attribute (one per time) among the out-of-bag objects.

The highest ranked attributes from each method as a function of classifier accuracy were compared, combined, and their individual impact on random forest classifiers was tested manually by forward selections of attributes (from the pre-identified subset of the most useful ones) until the classifier accuracy (i.e. completeness) reached a maximum or did not increase significantly. The tuning of random forest parameters (the numbers of trees and of attributes tested at each node) was automatically included in all of the optimisation procedures (because attribute changes imply new classifiers).

The attributes were selected in both initial (supervised) and final (semi-supervised) classification runs with the restrictions mentioned in the first paragraph of item (E) in Sect. 3.1.2. For the final classification, 40 attributes were selected from the union of the attributes employed in the five random forest classifiers detailed in Sect. 3.2. Except for a few astrometric parameters, most of the attributes described features in the G, $G_{\rm BP}$, and $G_{\rm RP}$ photometric time series (and from their combinations) in terms of statistical values. For a homogeneous treatment of all sources, considering that about half of the sources (or more, depending on the variability type) had fewer than 20 FoV transits in the G band, time series were not modelled by Fourier series decompositions, thus attributes did not include the characteristics that are typically employed to identify periodic variable objects, such as periods and comparisons of amplitudes and phases of different harmonics. All of the attributes related to the photometry were computed on cleaned photometric time series (Holl et al. 2018). The definitions of each attribute and of the subsets of attributes for each classifier are presented in Sect. 7.3.3.3 of Eyer et al. (2018).

As mentioned in Sect. 2, the calculation of astrometric attributes (PARALLAX, PROPER_MOTION, PROPER_MOTION_ERROR_TO_VALUE_RATIO) could only make use of a preliminary (not published) astrometric solution. The quasar-specific attributes (LOG_QSO_VAR, LOG_NONQSO_VAR, NONQSO_PROB) were computed from a parameterised quasar variance model (Butler & Bloom 2011) with parameter values that were determined from a preliminary per-FoV G-band photometry.

3.2. Classifier

The machine-learning algorithm employed for this classification was random forest (Breiman 2001), which averages the results of multiple decision trees with randomness in the selection of the data for each tree and in the selection of attributes at each node, typically leading to high accuracy (low bias and variance), robustness to noisy and correlated attributes, reduced susceptibility to overfitting, and other advantages (e.g. see Chap. 15 of Hastie et al. 2009).

The ESA *Gaia* archive[1] includes two tables that describe the global characteristics of a classifier:

– `gaiadr2.vari_classifier_definition`: this includes the classifier name and a brief description of the classifier (when multiple classification results are published for the same sources in the future);
– `gaiadr2.vari_classifier_class_definition`: this includes labels and brief descriptions of the variability classes of the classified objects.

3.2.1. Multi-stage classification

As a consequence of the numerous classes in the training set (Sect. 3.1.2), a single classifier is not an optimal choice in general because the multiple features necessary to recognise all of the details of all classes and subclasses at the same time unavoidably dilute the training data in attribute space and propagate the reduced statistical significance to the classification results. Moreover, failures in main classifications or subtle subclassifications are equally penalised, so that sources that belong to challenging subtypes can easily be assigned to completely different parent classes. On the other hand, a hierarchical organisation of classifiers dedicated to solve simpler problems in sequence can return more trustworthy results because (i) fewer attributes are needed to split a smaller number of (super)classes, further protecting from overfitting, (ii) each (super)class representation is statistically more significant, and (iii) each classifier model is easier to understand, which is an important advantage in the interpretation of results. However, the cumulated accuracy of a sequence of unavoidably imperfect classifiers decreases after each stage, so that multi-stage classifiers should only split class groups with very little confusion among them, and the high accuracy of classifiers is especially important at the nodes preceding the targeted classes in order to limit false positives and negatives as much as possible, which irreversibly propagate in the wrong branches of the tree.

Dedicated classifiers were configured at the nodes of a multi-stage decision tree and controlled the levels at which (subsets of) classes could be compared and then separated. Classes were grouped because of similar physical origins or common observational challenges. For example, low-amplitude variables are likely associated with noisy attributes and it is worth separating their classification from the one of other objects because a classifier trained with sources of similar quality is better suited to recognise their features (Long et al. 2012). Moreover, it is often easier to identify the common characteristics of a superclass (e.g. RR Lyrae) than its exact subclass (such as a double-mode RR Lyrae subtype, which typically requires more measurements than are available for most sources in *Gaia* DR2). The multi-stage top-down approach enables distinguishing groups of classes that share similar global features and to progressively split them by employing more detailed information, until the targeted (sub)class levels are reached. The classification probabilities of sources belonging to specific (groups of) classes are distributed from the top to the bottom levels according to the probabilities of such (groups of) classes to be assigned by the classifiers in each node (classifier probabilities express the fractions of the previous node probabilities to be passed to the next stage for the same groups of classes; see also Sect. 4.5.3 in Eyer et al. 2017).

In the multi-stage classification employed here, classification results followed the paths of the nodes with the highest probabilities along the same branches (i.e., lower-level nodes did not override the classifications of higher-level nodes): results were associated with the highest classifier probabilities at the end nodes of each branch, although they did not necessarily correspond to the highest probabilities of the nodes at the end of all branches. This decision followed from the following considerations:

1. Classifier probabilities do not express real (calibrated) probabilities (see Sect. 3.4) and the meaning of their values depends strongly on each class because of the specific sources in the training set, the relative number of trained sources per class, the setup of the multi-stage tree of classifiers, and the different number of nodes in different branches (lower optimal probabilities are expected from longer sequences of classifiers), among other factors (including the classification algorithm). Thus, the products of such probabilities (after each node) should not be interpreted as real probabilities either.
2. Classifiers that are higher in the hierarchy are deemed more trustworthy than those at lower levels because the required information for detailed classification might not be available for some sources (e.g. if poorly sampled), because key attributes for identifying specific classes might not be used for every node preceding these classes, and because high accuracy is typically required for higher-level classifiers.

This multi-stage classification guaranteed that insufficient representation or information for subtype identification did not jeopardise the parent-class membership, although the final (sub)type identification could be mistaken.

3.2.2. Implementation

The multi-stage tree with the respective classifiers after the semi-supervised phase described in item (E) of Sect. 3.1.2 are depicted in Fig. 3, where numbered shaded regions (from 1 to 5) and their contents denote classifiers and the classes they are meant to classify, respectively. These five classifiers were assessed through confusion matrices (which verify the classifications of training objects, accounting for the known and classified sources in rows and columns of a matrix, respectively), which were estimated from the out-of-bag sources in random forest. The random forest method was applied to the classifiers employing the implementation of Weka (Frank et al. 2016). The configuration parameters and the subsets of classification attributes employed for each classifier were listed in Sect. 7.3.3.4 of Eyer et al. (2018), while the most useful features for each classifier are mentioned in the following items (for the definitions of attributes, see Sect. 7.3.3.3 of Eyer et al. 2018).

– Classifier 1. The variable objects of interest were separated from constant and typically low-amplitude (percent level) variables. Because the calibration and the uncertainties of epoch photometry (Evans et al. 2018; Riello et al. 2018) were not yet optimal (Busso et al. 2018) and the presence of spurious measurements rendered variability detection with standard metrics and theoretical expectations impractical (see Sect. 7.2.3.4 in Eyer et al. 2018), a data-driven approach with a classifier provided a viable alternative to infer source variability levels. As the perceived constancy of objects depended on the precision of measurements, sources close to the variability detection limit could be classified as either constant or variable. In order to further distinguish this group of sources (and reduce the contamination of other classes), a low-amplitude class was introduced in addition to those for constant and other variable objects. The confusion matrix of this classifier is shown in Fig. 4a. Constant objects were recovered with very high completeness (99%) and with

# Obj. / Class		CONSTANT	LOW_AMPL	OTHER_VAR	
14684	CONSTANT	99	1	1	%
3885	LOW_AMPL	8	79	13	%
14999	OTHER_VAR	1	3	95	%
	Contamination	4	16	4	%

(a)

# Obj. / Class		ACV_ACYG_BCEP_GCAS_SPB_SXARI	DSCT_GDOR	ELL	OSARG	SOLARLIKE_FLARES_ROT	
363	ACV_ACYG_BCEP_GCAS_SPB_SXARI	96	3			1	%
866	DSCT_GDOR	1	93			6	%
397	ELL	4	2	84	3	6	%
996	OSARG				98	2	%
1247	SOLARLIKE_FLARES_ROT		4	1		95	%
	Contamination	8	7	6	1	7	%

(b)

Fig. 4. Confusion matrices of classifiers 1 (*panel a*) and 2 (*panel b*), as denoted in Fig. 3. The classifications of training objects (in rows) are compared with classifier results (in columns), which are estimated from the out-of-bag sources in random forest. Given the amount of true positives (TP), false positives (FP), and false negatives (FN), the completeness [TP/(TP+FN)] and contamination [FP/(TP+FP)] rates, expressed as rounded per-cent values, appear in the diagonal (in black) and the bottom row (in red), respectively, while the numbers of training objects per class are listed in blue on the left-hand side of each matrix. Rounded rates imply that not all rows sum to 100%. Rates below 0.5% are not shown to facilitate the reading of the most relevant parts. Darker shaded squares are used to highlight higher occurrence rates.

# Obj. / Class		BLAP	CEP_ACEP_T2CEP	CV	DSCT_SXPHE	ECL	MIRA_SR	QSO	RRAB_RRC_RRD_ARRD	RS	
10	BLAP				50				50		%
711	CEP_ACEP_T2CEP		68		1	5	2	2	19	4	%
518	CV			74	4	10	1	6	6	1	%
1326	DSCT_SXPHE			1	81	5		2	11	1	%
3861	ECL				1	94		1	3	1	%
1945	MIRA_SR						98			1	%
1996	QSO					1		97	1		%
4108	RRAB_RRC_RRD_ARRD				2	6		1	90	1	%
500	RS		3	1	4	11	4	1	3	74	%
	Contamination	–	11	10	13	12	3	7	11	24	%

Fig. 5. Same as Fig. 4, but for classifier 3 (as denoted in Fig. 3).

contamination rates of 8 and 1% from the low-amplitude and other variables, respectively. The lower weight of low-amplitude variables, implicitly assigned by the smaller number of representatives with respect to the other two classes, was designed to reduce contamination in the class of the other variables without seriously competing against the class containing the targeted variables. The loss of 3% of the other variables to low-amplitude objects was expected because a small fraction of low-amplitude objects was not removed from the other variables (where 3% of the objects had a G range of less than 0.03 mag, 97% of which were δ Scuti/SX Phoenicis stars, in addition to a handful of RS Canum Venaticorum-type stars, because the sub-classifications of these stars based on amplitude was often not available in the literature). The most useful attributes of this classifier were REDUCED_CHI2_G, DENOISED_UNBIASED_UNWEIGHTED_VARIANCE, NORMALIZED_CHI_SQUARE_EXCESS, DURATION, RANGE_G, MAD_G, STETSON_G, G_VS_TIME_MEDIAN_ABS_SLOPE, the colours from the three *Gaia* bands, and the median magnitudes.

– Classifier 2. Although the classifications of time series that belonged to classes with typically low magnitude ranges in the G band (dominated by percent level variations, over 80% of which were below 0.1 mag) were not published, the confusion matrix of the classifier is still shown in Fig. 4b to highlight the composition of low-amplitude classes, which included a set of δ Scuti stars (merged with γ Doradus stars and δ Scuti-γ Doradus hybrids) that exhibited low-amplitude variations. Some classes (such as γ Cassiopeiae) are listed among low-amplitude variables because no high-amplitude sample was available in the training set. Low-amplitude γ Cassiopeiae stars were merged with pulsating and other multi-periodic types because eruptions and apparent irregularities due to multi-periodicity (in addition to similar colours and amplitudes) can cause their time series to look rather similar, with the sampling of *Gaia* DR2. The most useful attributes included MEAN_G, PARALLAX, BP_MINUS_RP_COLOUR, and MEAN_BP.

– Classifier 3. The key identifications of the published high-amplitude pulsating variables were returned by this classifier. Focusing on the classes relevant to *Gaia* DR2, the confusion matrix in Fig. 5 shows that the RR Lyrae stars and long-period variables (Miras and semiregulars) were identified with an accuracy of at least 90%, while the less well represented δ Scuti/SX Phoenicis and Cepheid classes reached a completeness of 81 and 68%, respectively. The overall contamination rates of these four groups of classes were limited to 3–13%. The main sources of contamination of RR Lyrae classifications (for which the RR Lyrae class was also the most common misclassification) were represented by Cepheids, δ Scuti/SX Phoenicis stars, and eclipsing binaries, as expected from a classifier that did not use Fourier parameters as attributes. The blue large-amplitude pulsators (BLAP, Pietrukowicz et al. 2017) were not sufficiently representative in the training set, and the confusion matrix confirmed that they had no chance of detection, hence it was decided to merge the BLAP classifications with those of δ Scuti/SX Phoenicis stars, where potential BLAP candidates (if any) were most likely to be found. Some of the attributes that were particularly useful in identifying the published classes in those trained for this classifier were as follows: G_VS_TIME_MAX_SLOPE and NORMALIZED_CHI_SQUARE_EXCESS for Cepheids; RANGE_G and DENOISED_UNBIASED_UNWEIGHTED_VARIANCE for δ Scuti/SX Phoenicis stars; BP_MINUS_RP_COLOUR, MEDIAN_RANGE_HALFDAY_

# Obj. / Class		ARRD	RRAB	RRC	RRD	
20	ARRD		20	80		%
2922	RRAB		97	3		%
969	RRC		12	86	2	%
197	RRD		35	57	8	%
Contamination		100	6	21	64	%

(a)

# Obj. / Class		ACEP	CEP	T2CEP	
99	ACEP	73	16	11	%
455	CEP	3	94	3	%
157	T2CEP	11	10	80	%
Contamination		29	7	17	%

(b)

Fig. 6. Same as Fig. 4, but for classifiers 4 (*panel a*) and 5 (*panel b*), as denoted in Fig. 3.

TO_ALL, and G_MINUS_RP_COLOUR for Miras/Semi-regulars; SKEWNESS_G and MEDIAN_ABS_SLOPE_ONEDAY for RR Lyrae stars.

– Classifier 4. A sub-classification of RR Lyrae stars into fundamental mode (RRAB), first-overtone (RRC), and (anomalous) double-mode types (A/RRD) was attempted, and the confusion matrix in Fig. 6a clearly shows that (anomalous) double-mode RR Lyrae types were almost always confused with the much more numerous fundamental mode and first-overtone ones, of which the former was identified with high completeness and limited contamination rates (97 and 6%, respectively). The most useful attributes were DENOISED_UNBIASED_UNWEIGHTED_KURTOSIS_MOMENT, TRIMMED_RANGE_G, and G_VS_TIME_MAX_SLOPE.
– Classifier 5. A sub-classification of Cepheids was also attempted, and the confusion matrix shown in Fig. 6b indicates decreasing levels of completeness, accompanied by increasing contamination rates, for classical, type-II, and anomalous Cepheids (in this order). The best performance was achieved by the classical Cepheids with 94 and 7% completeness and contamination, respectively. The classification attributes that proved to be the most useful included MEDIAN_RP, LOG_NONQSO_VAR, and BP_MINUS_RP_COLOUR.

Classifier validations from the confusion matrices presented in Figs. 4–6 included all classification probabilities and were limited to training objects, these assessments therefore depended on the choice of sources, classes, and their relative representation, and they were naturally biased (by definition) against untrained sources of the same classes that looked different for some reason (e.g. in case the effects of interstellar extinction were not fully accounted for in the training set). These validations gave valuable indications for building classifier models, but classification results should be assessed independently of these preliminary estimates.

This multi-stage classifier was applied to sources with at least two FoV transits in the G band and with the trimmed range (from the 5th to the 95th percentile) of epoch photometry in the G band greater than 0.1 mag. The results were assessed as a function of classification probability, and excessive numbers of weak candidates were excluded by setting minimum probability thresholds, if needed. In particular, class labels related to candidates that were not affected by probability thresholds were ACEP, ARRD, and RRD, while minimum classification probabilities of 0.3, 0.4, 0.4, 0.58, 0.6, and 0.6 were applied to candidates of classes labelled as MIRA_SR, CEP, T2CEP, DSCT_SXPHE (after summing the probabilities of BLAP candidates), RRAB, and RRC, respectively. This subset of classification results was then filtered by validation classifiers (unpublished), as described in Sect. 3.3.

3.3. Validation

The first selection of classification results, described in Sect. 3.2.2, still included significant contamination (just by looking at the number of candidates for each class). In order to alleviate the presence of contaminants in an automated way, new dedicated (validation) classifiers were built to help separate false positives from true positives. This binary classification was pursued with random forest classifiers, one for each of the three groups of classes labelled as CEP_ACEP_T2CEP, DSCT_SXPHE, and RRAB_RRC_RRD_ARRD. The training sets employed a similar number of true positives and false positives, which were sampled from the classified *Gaia* sources in common with the 750 thousand cross-matched objects of known types in the literature (defined in Sect. 3.1.1). Except for the DSCT_SXPHE candidates (which were classified after training with all of the δ Scuti/SX Phoenicis stars cross matched with the literature), the validation classifiers were trained with cross-matched objects that were not used by the preceding classification stage (Sect. 3.2). Similar to the other classifiers, the validation training sources were sampled for a representative distribution in the sky, and their attributes were selected as described in Sect. 3.1.3, with the consequent optimisation of random forest parameters. Validation classifiers were applied to the preliminary selection of classification results (Sect. 3.2.2) and further classified the candidates of the three superclasses mentioned above as true positive or false positive (with a validation classification probability threshold of 0.5). Only the true-positive identifications were published in the variability classification table of *Gaia* DR2 and are described in Sect. 4.

The completeness rates of true positives were 94, 98, and 91%, with corresponding contamination rates of 3, 3, and 9%, for the CEP_ACEP_T2CEP, DSCT_SXPHE, and the RRAB_RRC_RRD_ARRD superclasses, respectively. Thus, contamination was reduced, at the cost of 2–9% of further reduction in completeness, with the greatest loss of true positives for the RR Lyrae candidates.

In the special case of MIRA_SR candidates, the training sources of long-period variables were not considered sufficiently representative to further validate the preliminary results. The published classifications therefore reflect the selection of the most likely candidates performed by the dedicated SOS module (Mowlavi et al. 2018).

Known misclassified objects from the literature were not removed from the classification results in order to preserve the consistency of the results and prevent the appearance of cross-match footprints, or statistical studies would face additional challenges to distinguish real from artificial features. Moreover, not all of the classifications available in the literature are necessarily correct, and in some cases, the *Gaia* data provide the additional information that can lead to an improved judgement.

3.4. Classification score

The classification scores and the corresponding class labels assigned to the classified variable sources are stored in the `vari_classifier_result` table of the ESA *Gaia* archive[1] under the field names `best_class_score` and `best_class_name`, respectively. The classifier score is a numerical quantity between 0 and 1 that expresses a linear

transformation of the classification probabilities as follows:

$$\texttt{best_class_score} = (P_{\rm class} - P_{\rm min,\,class})/(1 - P_{\rm min,\,class}),$$

where $P_{\rm class}$ denotes the classification probability (in the range from 0 to 1) of the best class from the multi-stage classifier (Sect. 3.2) and $P_{\rm min,class}$ refers to the minimum probability thresholds as a function of class, listed in the last paragraph of Sect. 3.2.2.

The motivation for this transformation followed from the attempt of assigning similar scores to classifications with similar reliability, considering that the probabilities of the weakest candidates depended on class. This simplistic approach should not be interpreted as returning true (calibrated) probabilities (e.g. Richards et al. 2012), which were not pursued for this data release. The classifier "probabilities" do not represent the real probabilities of classifications to be true positives because of their dependence on the classifier method and the training set (such as the number of classes, the number of sources for each class, the extent of representation of the selected sources, and the setup of the multi-stage tree, as mentioned in Sect. 3.2.1).

4. Results

The classifications of pulsating variable stars of Cepheid, Mira/semiregular, δ Scuti/SX Phoenicis, and RR Lyrae types, with light-variation ranges greater than 0.1 mag in the G band (increased to 0.2 mag for long-period variables), are published in the table `vari_classifier_result` of the ESA *Gaia* archive[1], which includes the source identifier, the classifier name (in case of multiple independent classifiers in the future), the class label, and the associated classification score.

An overview of the classification results that satisfy the astrometric and photometric requirements for an observational Hertzsprung–Russell diagram (see Appendix B of Gaia Collaboration 2019) is presented in Fig. 7. The effect of extinction is visible for several classes (in particular, long-period variables, classical Cepheids, and fundamental mode and first-overtone RR Lyrae stars). In addition, faint outliers, with respect to the loci of other candidates of the same class, identify contaminants for most of the classes (assuming accurate parallax values), which are discussed in more detail for each variability type in this section.

Samples of light curves for each class are shown in the summary article of variables in *Gaia* DR2 (Holl et al. 2018) and in the related SOS articles (Clementini et al. 2019; Mowlavi et al. 2018); they are therefore not reproduced here. The goal of this section is to present the global properties of the classified candidates for each class group (colour-magnitude diagrams, distributions in the sky versus magnitude, and classification score), outline specific features of the candidates unconfirmed or reclassified by SOS, and finally compare the classifications of a subset of sources with those that are known in the literature.

The comparison of *Gaia* classifications with the literature can be useful in the assessment of some aspects related to completeness and contamination. However, results in the literature are not exempt from misclassifications, and they depend relevantly on the observational properties (e.g. photometric bands, sky location, magnitude limits, signal-to-noise ratio, astrometric resolution, availability of spectroscopic and astrometric information, and number of epochs and their sampling), on the classification processing (e.g. reliability level, targeted completeness, and purity levels), and on the variability types (some of which may elude detection or proper classification depending on observation times or stellar evolution) and their occurrence

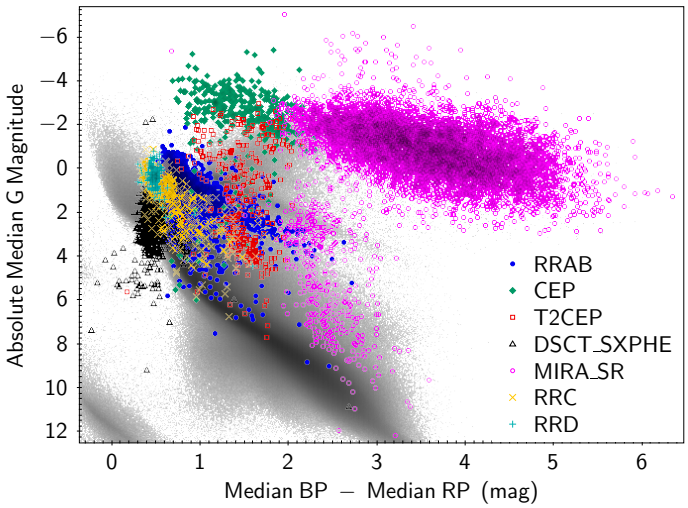


Fig. 7. Colour-absolute magnitude diagram (as absolute median G-band magnitude vs. median $G_{\rm BP}$− median $G_{\rm RP}$) of a selection of classification candidates: fundamental mode (RRAB, blue dots), first-overtone (RRC, orange "×" marks), and double-mode (RRD, cyan "+" marks) RR Lyrae types, classical (CEP, green rhombi) and type-II (T2CEP, red squares) Cepheids, δ Scuti/SX Phoenicis types (DSCT_SXPHE, black triangles), and Mira/semiregular types (MIRA_SR, magenta circles). Faint outliers typically denote contaminating objects, which in the case of long-period variables, are represented mostly by young stellar objects. All of these sources satisfy the conditions listed in Appendix B of Gaia Collaboration (2019), among which a relative parallax precision better than 20%, with no correction for interstellar or circumstellar extinction or reddening. The background points in grey act as a reference of objects within a radius of 1 kpc from the Sun.

relative to each other (i.e. in the literature versus in reality). Comparisons with cross-matched sources of known types from the literature are therefore influenced by the different survey features and criteria. On the other hand, no survey provides a perfect reference in all circumstances, and the diversity of a multitude of literature catalogues can help overcome some of the limitations of single surveys.

We here compare the *Gaia* classifications with a subset of 494 thousand sources that have a G-band range greater than 0.1 mag of the 750 thousand cross-matched objects defined in Sect. 3.1.1, in addition to new *Gaia* source cross matches with the active galactic nuclei identified in the mid-infrared using the final catalogue release of the Wide-field Infrared Survey Explorer (AllWISEAGN; Secrest et al. 2015) and the quasars listed in the third release of the Large Quasar Astrometric Catalog (LQAC; Souchay et al. 2015). Despite the variety of classes considered, 94% of the known objects were represented by constant, eclipsing binary, and RR Lyrae stars. From this set of cross-matched sources, we expect biases in the apparent rates of completeness and contamination (especially for the latter as it relies on the inclusion of all contaminating classes, on the adoption of realistic relative class proportions, and on a negligible contribution from unexpected sources), in addition to the optimistic selection bias from using classifications from the literatures (part of which might be easier to identify). Nevertheless, we present completeness and contamination rates for each class (or subclass, when applicable) and as a function of minimum classification score in order to show trends (such as the most or least pure subclass) and give insights into likely sources of contamination.

As mentioned in Sect. 3.2.2, four (super)classes constitute the primary results of relevance. Subclassifications are provided,

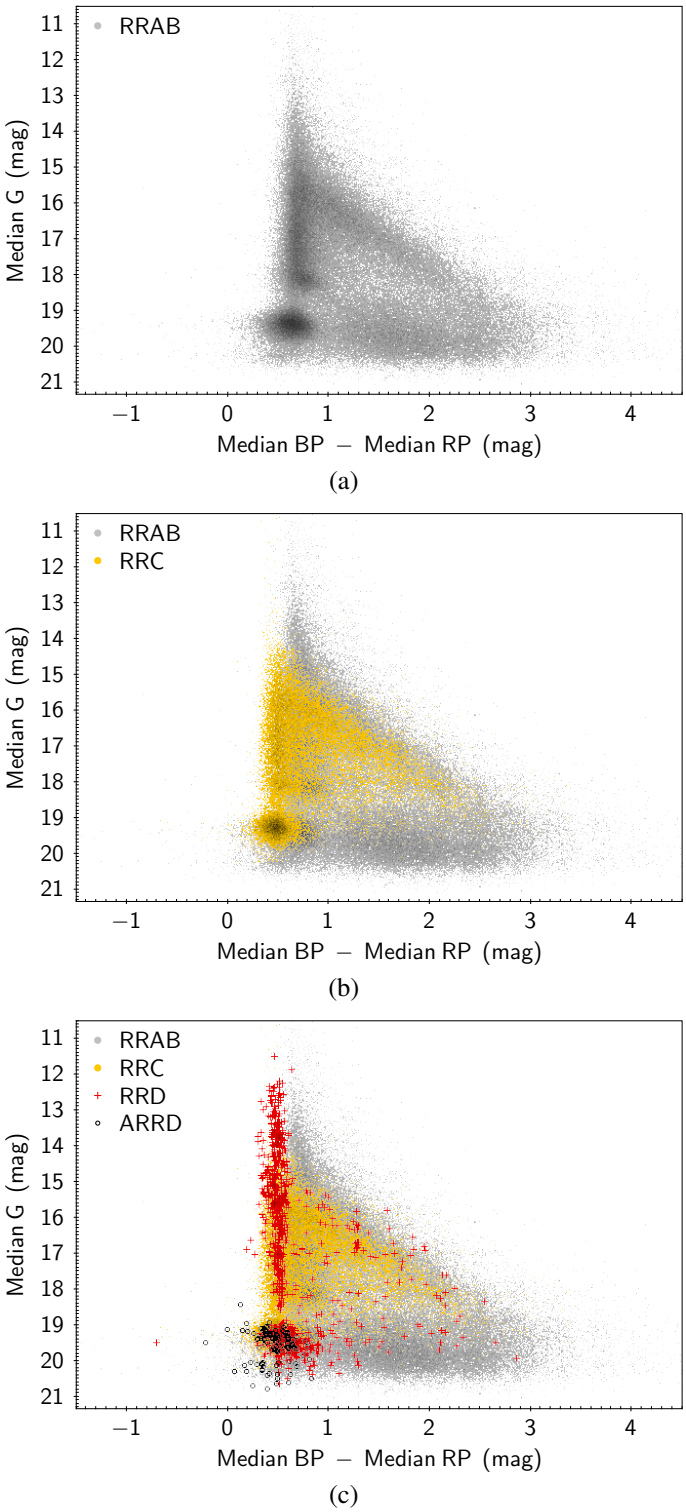


Fig. 8. Colour-magnitude diagram (as median G-band magnitude vs. median $G_{\rm BP}$- median $G_{\rm RP}$) of the RR Lyrae classifications. The three panels show the fundamental mode (labelled RRAB, grey dots), first-overtone (labelled RRC, orange dots), and (anomalous) double-mode (labelled ARRD/RRD, black circles/red crosses) RR Lyrae subtypes, cumulatively, in *panels a–c*, respectively. See text for the explanation of the main features.

but their use should be limited to the identification of stars that belong to the most common subtypes (for reduced contamination levels) or simply to the membership of their parent classes.

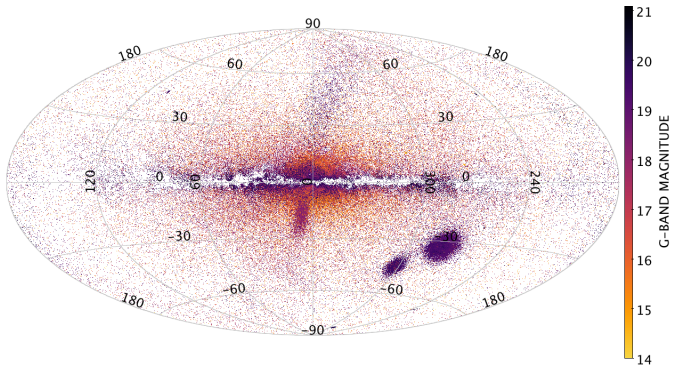


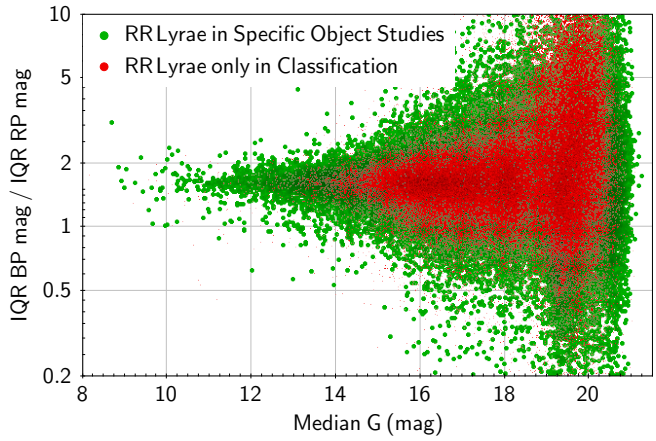
Fig. 9. Distribution of the RR Lyrae classifications (all subtypes) in the sky (Galactic coordinates in degrees) colour-coded by the median G-band magnitude as indicated in the legend on the right-hand side (values beyond the legend range share the colour of the closest value in the legend).

4.1. RR Lyrae stars

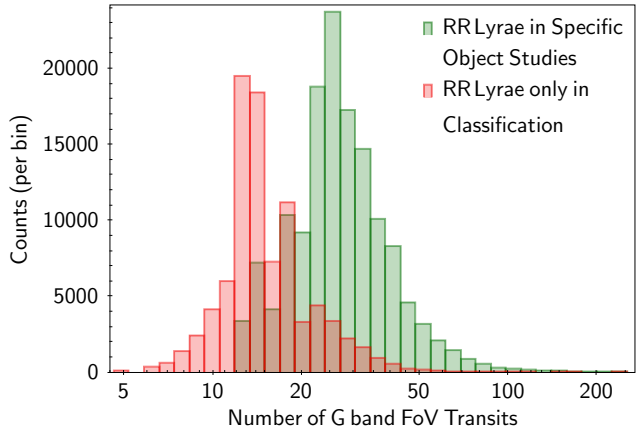
The RR Lyrae classifications include 195 780 candidates, further subclassified into 162 469 fundamental mode (RRAB), 32 370 first-overtone (RRC), 834 double-mode (RRD), and 107 anomalous double-mode (ARRD) subtypes. Their distribution in a colour-magnitude diagram is shown in Fig. 8 in terms of median G magnitude versus $G_{\rm BP} - G_{\rm RP}$ colour. The unreddened colour of RRAB candidates is centred between $G_{\rm BP} - G_{\rm RP}$ of 0.6 and 0.7 mag, as inferred from the sources in the Galactic halo that form the vertical structure observed in Fig. 8a. The slightly reddened clump around $G \approx 18$–18.4 mag and $G_{\rm BP} - G_{\rm RP} \approx 0.8$ mag is associated with about 3 thousand candidates in the Sagittarius dwarf spheroidal galaxy, which is visible just below the Galactic bulge in Fig. 9. The faintest unreddened clump at about $G \approx 18.8$–20 mag is primarily due to almost 20 thousand RRAB candidates in the Magellanic Clouds and to classifications that belong to the Sagittarius stream above the Galactic bulge. Some of the fainter objects are still related to the Magellanic Clouds and others to dwarf spheroidal galaxies (Clementini et al. 2019). The diagonal branch in Fig. 8a is due to candidates that are reddened and extinguished mostly by the Galactic dust in the disc and bulge regions, with redder and fainter candidates located closer to the Galactic equator. Objects between the vertical and diagonal features, including the horizontal overdensity at the faint end with $G_{\rm BP} - G_{\rm RP} > 1$ mag, are dominated by 20–25% of likely misclassified objects in the Galactic disc or bulge.

The RRC candidates show similar structures to those of RRAB candidates in Fig. 8b, with an offset of almost 0.2 mag bluer on average than RRAB candidates, a reduced contamination at the faint end, and about 5 and 0.5 thousand candidates in the regions of the Magellanic Clouds and Sagittarius dwarf spheroidal galaxy, respectively. About 20% of the RRD candidates shown in Fig. 8c are in the Magellanic Clouds, and the others are distributed mostly in the halo. More than half (about 60%) of the ARRD candidates are in the Magellanic Clouds because the training set included only OGLE-IV samples from this region; other ARRD candidates are scattered across the sky.

The processing of the RR Lyrae star prototype was particularly unfortunate in *Gaia* DR2: the published values of its mean photometry and parallax were inaccurate (Arenou et al. 2018; Gaia Collaboration 2018) and it was missed also from the all-sky classification of variables because the sampling of the signal made the time-series statistics (without time-series mod-



(a)



(b)

Fig. 10. *Panel a*: distribution of the ratios of the interquartile range (IQR) in the G_{BP} versus G_{RP} bands vs. the median G-band magnitude for the RR Lyrae candidates in the SOS (green dots; Clementini et al. 2019) vs. those present only in the classification results (red dots). *Panel b*: distribution of the number of G-band FoV transits for each source for the RR Lyrae candidates in the SOS (green bars) vs. those present only in the classification results (red bars).

elling) similar to eclipsing binary variations. Only 10 of the 26 valid measurements in the G band were in the faint half of the magnitude range, leading to a positive skewness of the magnitude distribution (typical of eclipsing binaries), and the random forest probability of this object to belong to the eclipsing binary class became slightly higher than the one of an RRAB type (0.5 versus 0.4, respectively). The SOS module for RR Lyrae stars received this object correctly classified by the independent classification run limited to sources with at least 20 FoV transits in the G band (which could take advantage of the periodicity information), but eventually it discarded this candidate because a key Fourier parameter was not sufficiently accurate (Clementini et al. 2019).

The RR Lyrae candidates with at least 12 FoV transits in the G band were considered for validation by the dedicated SOS module with Fourier modelling, as described in Clementini et al. (2019), although this lower limit was not always sufficient to guarantee the confirmation of the related classifications. In total, 88 120 RR Lyrae classifications were not confirmed in SOS. Because stellar pulsations are expected to exhibit larger variations in the G_{BP} than in the G_{RP} band, the distribution of the ratios of the interquartile range (IQR) in the G_{BP} versus

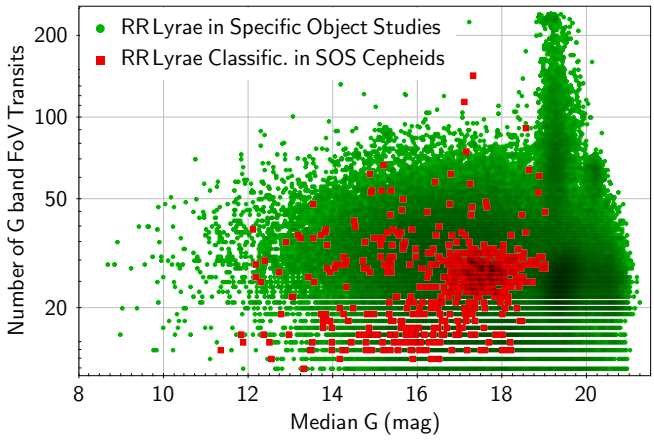


Fig. 11. RR Lyrae classifications (red squares) that were reclassified as Cepheids in the SOS (green dots; Clementini et al. 2019) as a function of median G-band magnitude and number of G-band FoV transits per source.

G_{RP} band is shown as a function of median G-band magnitude in Fig. 10a for the RR Lyrae candidates in SOS versus those present only in the classification results. Although the IQR is increasingly influenced by the photometric noise towards fainter magnitudes, 78 194 of the RR Lyrae classifications (89% of the candidates not confirmed in SOS) have IQR(G_{BP})/IQR(G_{RP}) >1. The distributions of the number of G-band FoV transits for each source for the SOS-confirmed versus unconfirmed RR Lyrae candidates are presented in Fig. 10b and are found to exhibit two distinct peaks, highlighting the importance of the number of observations in the SOS modelling of these objects and thus their confirmation process. A total of 618 RR Lyrae classifications are reclassified as Cepheids in SOS, as shown in Fig. 11 as a function of median G-band magnitude and number of G-band FoV transits for each source: 593 and 436 (96 and 71%) are labelled as RRAB and have classification scores lower than 0.5, respectively. Of the RR Lyrae stars that are misclassified as Cepheids, those that form a clump with median G of approximately 16.5–18.5 mag and with a similar number of observations amount to slightly more than half of the sample and are located in the region of the Magellanic Clouds. The other misclassified candidates are scattered across most of the sky, with a higher occurrence in the region of the Galactic bulge.

The distribution in the sky of the classification scores of RR Lyrae types is presented in Fig. 12, without particularly noticeable variations as a function of sky region. A comparison of the apparent completeness and contamination rates of RR Lyrae candidates of any score, employing sources that were cross matched with the literature, is shown in Figs. 13 (in counts) and 14 (in percentage), after excluding all training-set objects. Most of the confusion seems to be among the RR Lyrae subtypes, and RRAB and RRD represent the most and the least complete, respectively. Contamination rates are clearly underestimated, but they should not be over-interpreted because of the reasons mentioned in the beginning of Sect. 4. Non-RR Lyrae contaminating classes include, in order of relevance, Cepheids (CEP), eclipsing binaries (ECL), and quasars (QSO).

One source of unexpected contamination (not included in Figs. 13 and 14) includes galaxies, especially as the G-band line spread function fitting of extended objects might return different flux levels as the *Gaia* spacecraft scan angle rotates (S. Cheng & S. Koposov, NYC *Gaia* Sprint 2018[5]). The source identifiers

[5] http://gaia.lol/2018NYC.html

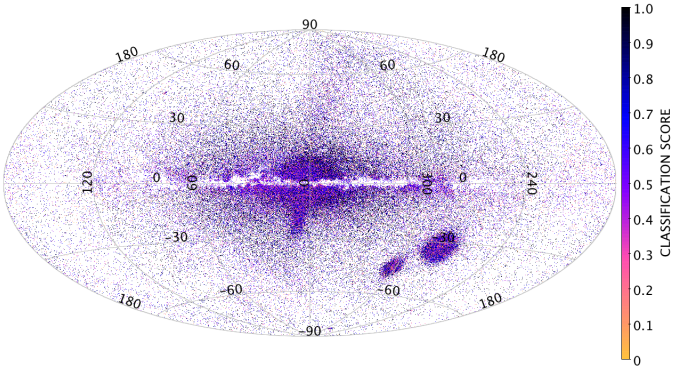


Fig. 12. Distribution of the RR Lyrae classifications (all subtypes) in the sky (Galactic coordinates in degrees) colour-coded by the classification score (see Sect. 3.4) as indicated in the legend on the right-hand side.

Literature

Gaia	ACEP	CEP	ECL	ELL	QSO	RRAB	RRC	RRD	RS	Contamination (%) by non−RR Lyrae
ARRD					4	9	18	4		11
RRAB	9	280	53	1	19	68805	532	441		1
RRC		19	37	1	2	462	12486	603	1	0
RRD		2	1		1	66	146	47		2

Classification score > 0

Fig. 13. Comparison between the number RR Lyrae classifications of any score per subtype (in rows) and the classes found for these objects in the literature (in columns), among the 494 thousand cross-matched sources with G-band range greater than 0.1 mag. Contamination rates (rounded per-cent values) with objects that do not belong to the RR Lyrae superclass are listed in the last column in red. Darker shaded squares indicate higher occurrences.

of 982 likely galaxies in the 140 784 RR Lyrae classifications confirmed by the dedicated SOS module are listed in Clementini et al. (2019).

Samples with higher completeness and lower contamination can often be selected by applying thresholds to classification scores and other quantities, such as brightness, variation amplitude, number of observations, and sky region. The dependence of the apparent completeness and contamination rates (derived from cross-matched sources) on minimum classification scores for the RR Lyrae subtypes is shown in Fig. 15, generally confirming the expected trend that higher score thresholds increase the completeness to contamination ratio more efficiently than lower score limits. In addition to the SOS module results described in Clementini et al. (2019) as part of the *Gaia* variability pipeline, an independent validation of the RR Lyrae classifications was performed with stars observed in selected *K2* fields of the *Kepler* space telescope (Molnár et al. 2018).

4.2. Cepheids

The Cepheid classifications include 8550 candidates, further subclassified into 6493 classical (or δ) Cepheids (CEP), 1743 type-II Cepheids (T2CEP), and 314 anomalous Cepheids (ACEP). Their distribution in a colour-magnitude diagram is shown in Fig. 16 in terms of median G magnitude versus $G_{\rm BP} - G_{\rm RP}$ colour for each subtype. Most of the classical Cepheids fainter than $G \approx 13$ mag (about 90% of all CEP classifications) are located in the region of the Magellanic Clouds, while the remaining (almost 10%) brighter candidates are found

Literature (%)

Gaia	ACEP	CEP	ECL	ELL	QSO	RRAB	RRC	RRD	RS	Contamination (%) by non−RR Lyrae
ARRD										11
RRAB	14	3				71	2	16		1
RRC							36	22		0
RRD								2		2

Classification score > 0

Fig. 14. Same as Fig. 13, but in terms of (rounded) per-cent values with respect to the total number of objects for each class in the literature. Columns do not sum to 100% because only rows relevant to RR Lyrae classifications are shown (and completeness is limited).

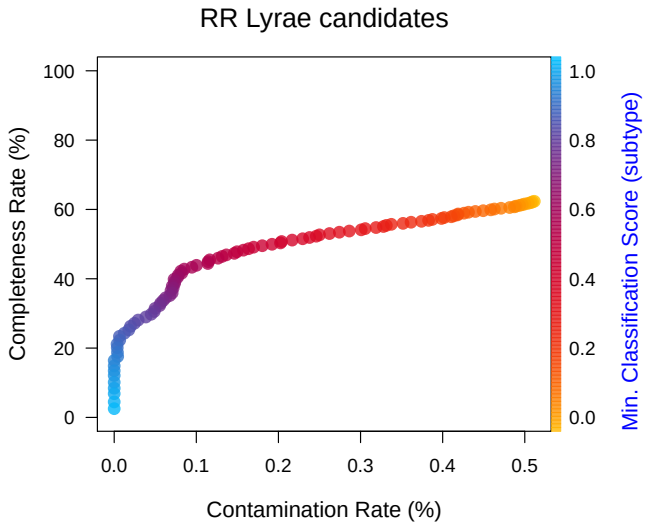


Fig. 15. Completeness vs. contamination rates of RR Lyrae classifications (all subtypes) as a function of minimum classification score (defined in Sect. 3.4 and colour-coded as indicated in the legend on the right-hand side).

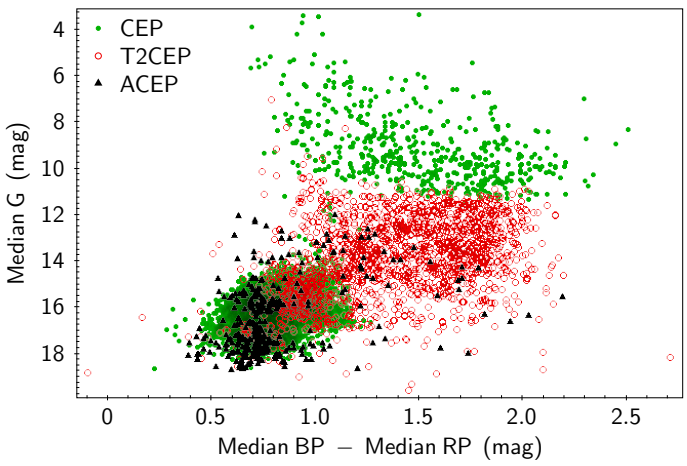


Fig. 16. Colour-magnitude diagram (as median G-band magnitude vs. median $G_{\rm BP}$− median $G_{\rm RP}$) of the Cepheid classifications, including the classical (labelled CEP, green dots), type-II (labelled T2CEP, red circles), and anomalous (labelled ACEP, black triangles) Cepheids.

primarily in correspondence of the Galactic disc, as shown also from the sky distribution in Fig. 17. The small number of classical Cepheids classified in the range of $G \approx 11$–13 mag is due to a training-set bias because about half of the classical

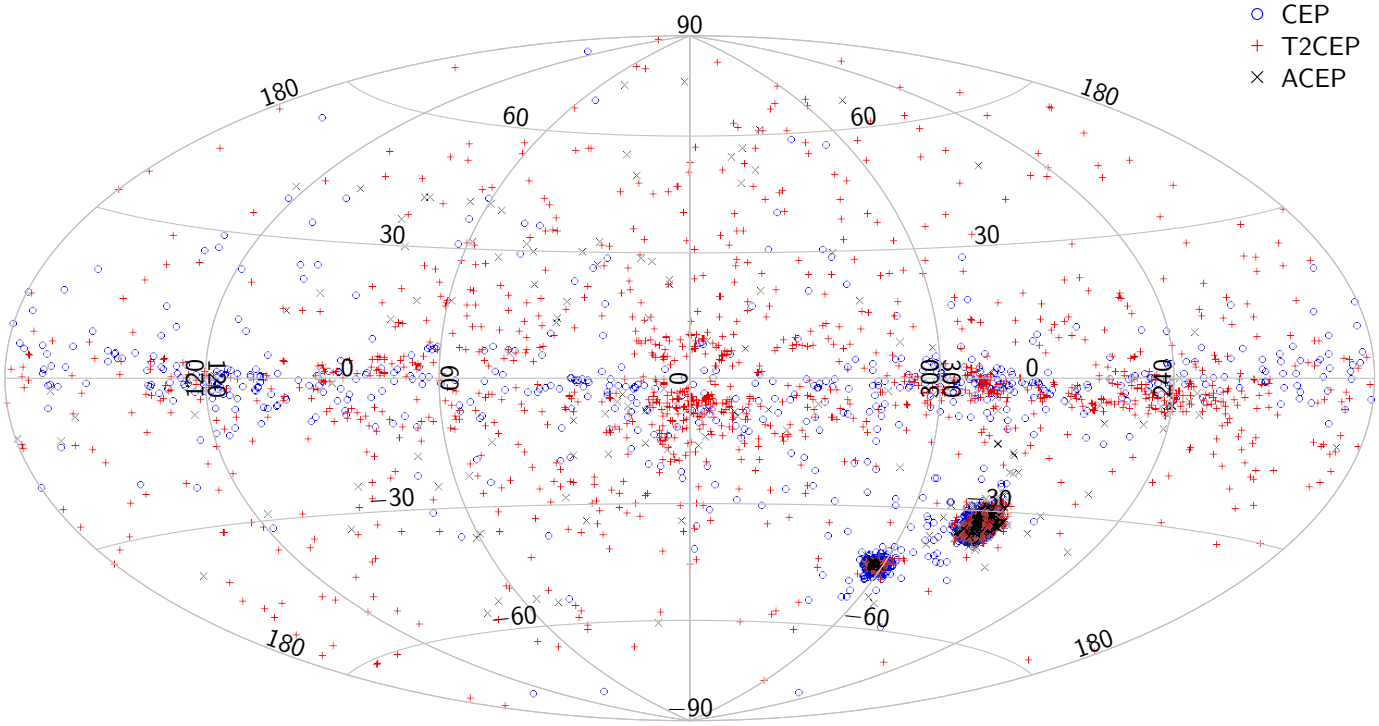


Fig. 17. Distribution of the Cepheid subtype classifications in the sky (Galactic coordinates in degrees), including the classical (CEP, blue circles), type-II (T2CEP, red "+" marks), and anomalous (ACEP, black "×" marks) Cepheids.

Cepheid sample was derived from HIPPARCOS classifications and the other half from the region of the Magellanic Clouds. Thus, the CEP classifications represent the magnitude ranges of the Magellanic Clouds and of the HIPPARCOS survey (with magnitude limit at $G \approx 11$ mag). Except for a handful of candidates, type-II Cepheid classifications are fainter than $G \approx 10$ mag, and almost one-third of them is associated with the Magellanic Clouds (typically bluer than $G_{\rm BP} - G_{\rm RP} \approx 1.5$ mag), while the rest is scattered in the Galactic halo, with slight overdensities below the Galactic bulge and in parts of the disc. The anomalous Cepheid classifications are fainter than $G \approx 12$ mag, and slightly more than half of them are found in the Magellanic Clouds (similar to the training-set composition and thus likely biased by it), while the other candidates are scattered across the whole sky.

The distribution in the sky of the median G-band magnitudes of the classified Cepheids, shown in Fig. 18, confirms the findings from Figs. 16 and 17 that except for the Magellanic Clouds, the Galactic disc includes mostly bright classical Cepheid classifications and that fainter candidates (near the disc or in the halo) are classified as type-II or anomalous Cepheids. The distribution of the classification scores of the Cepheid candidates in the sky is presented in Fig. 19 and suggests less certain identifications in the region of the Sagittarius dwarf spheroidal galaxy and in a some patches of the Galactic disc that are correlated with the overdensities of type-II Cepheid classifications.

Similar to the RR Lyrae classifications (Sect. 4.1), the Cepheid candidates with at least 12 FoV transits in the G band were considered for validation by the dedicated SOS module (Clementini et al. 2019) and a total of 1863 Cepheid classifications were not confirmed in SOS. The distribution of the IQR in the $G_{\rm BP}$ versus $G_{\rm RP}$ band is shown as a function of median G-band magnitude in Fig. 20 for the Cepheid candidates in SOS versus those present only in the classification results, and 1654 of the latter (89% of the candidates not confirmed in SOS) have IQR($G_{\rm BP}$)/IQR($G_{\rm RP}$) > 1. A total of 77 Cepheid classifications

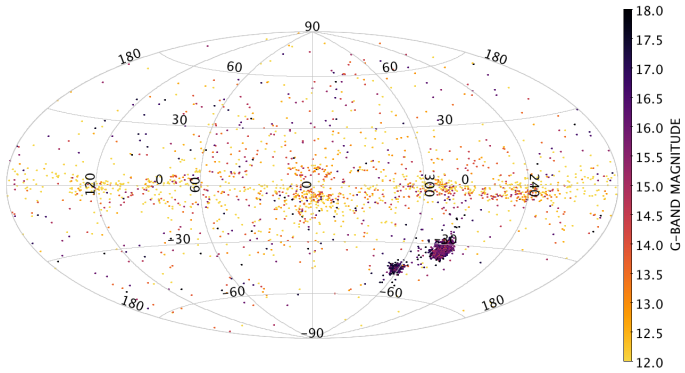


Fig. 18. Same as Fig. 9, but for Cepheid classifications (all subtypes).

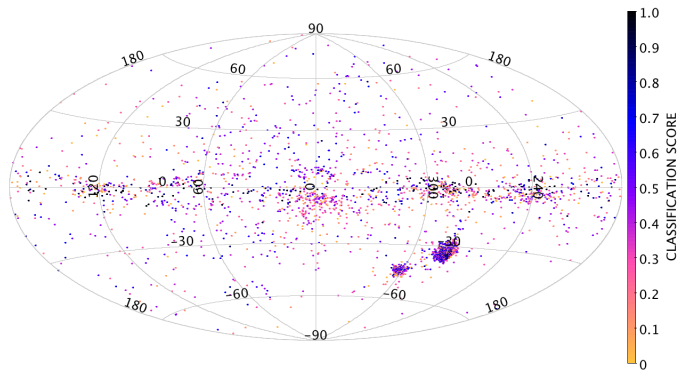


Fig. 19. Same as Fig. 12, but for Cepheid classifications (all subtypes).

are reclassified as RR Lyrae stars in SOS, as shown in Fig. 21 as a function of median G-band magnitude and number of G-band FoV transits per source: 40 and 72 of them (52 and 94%) are

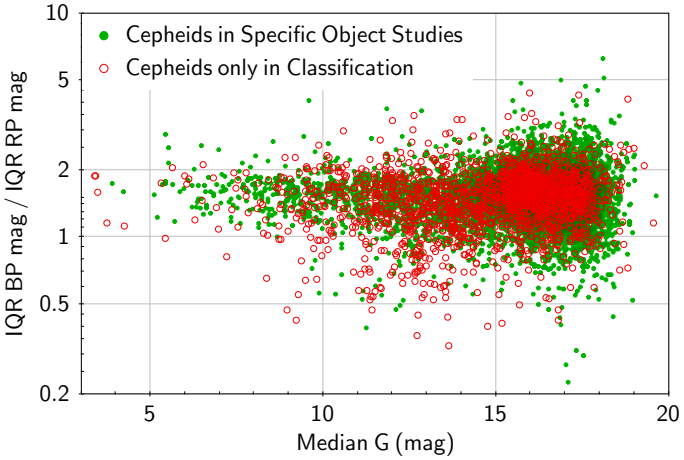


Fig. 20. Same as Fig. 10a, but for the Cepheid candidates in the SOS (green dots; Clementini et al. 2019) vs. those present only in the classification results (red circles).

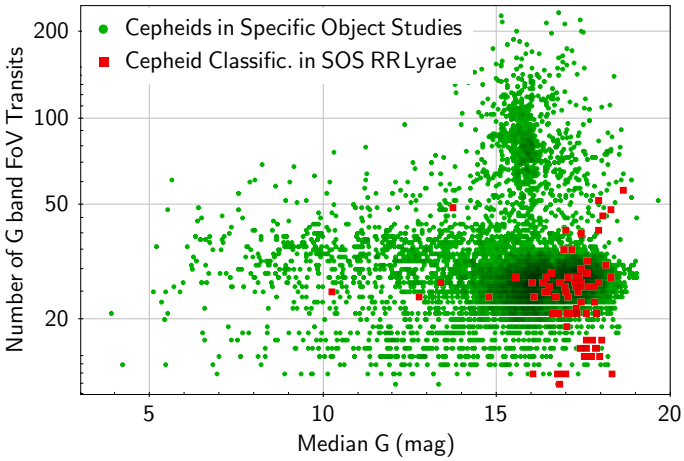


Fig. 21. Same as Fig. 11, but for the Cepheid classifications (red squares) that were reclassified as RR Lyrae stars in the SOS (green dots; Clementini et al. 2019).

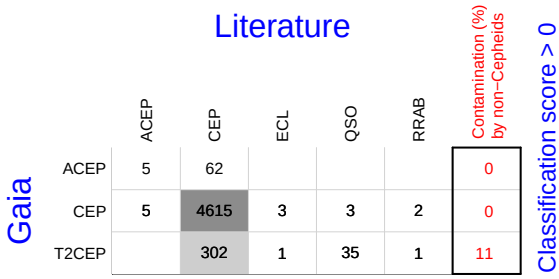


Fig. 22. Same as Fig. 13, but for Cepheid classifications.

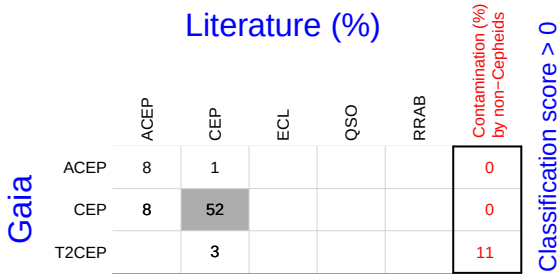


Fig. 23. Same as Fig. 14, but for Cepheid classifications.

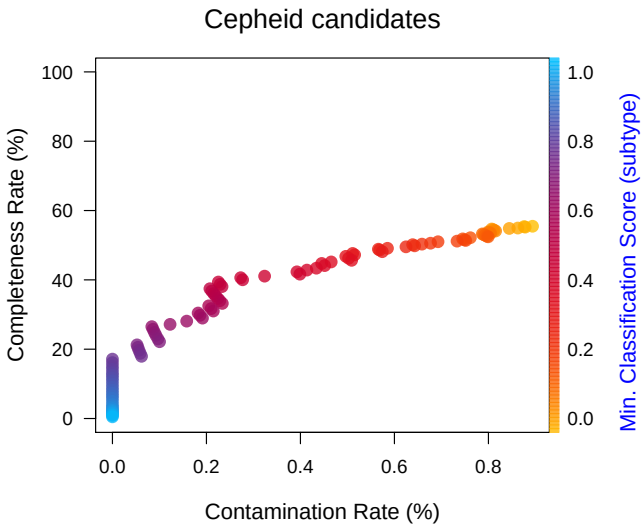


Fig. 24. Same as Fig. 15, but for Cepheid classifications.

labelled as ACEP and have classification scores lower than 0.5 (typically at the faint end), respectively. In the special case of the δ Cephei prototype, the latter is correctly identified as a classical Cepheid in the classification results, but it is absent from the SOS results because the source colours, the recovered period, and the related Fourier parameters were affected negatively by the limited sampling of this object.

A comparison of the apparent completeness and contamination rates of Cepheid candidates of any classification score, employing sources that were cross matched with the literature, is shown in Figs. 22 (in counts) and 23 (in percentage), after excluding all training-set objects (and all of the T2CEP cross-matched sources, as all of them were used in the training set). The completeness of the classical Cepheids, much better represented in the training set, is superior to the anomalous Cepheid classifications, as expected. Contamination rates are clearly underestimated, but they should not be over-interpreted because of the reasons mentioned in the beginning of Sect. 4. A small set of quasars (QSO) seems to be the primary source of contamination, especially for T2CEP classifications.

The dependence of the apparent completeness and contamination rates (derived from cross-matched sources) on minimum classification scores for the Cepheid subtypes is shown in Fig. 24. It generally confirms the expected trend that higher score thresholds increase the completeness-to-contamination ratio more efficiently than lower score limits.

In addition to the SOS module results described in Clementini et al. (2019) as part of the *Gaia* variability pipeline, an independent validation of a small sample of Cepheid classifications was performed with stars observed in selected *K2* fields of the *Kepler* space telescope (Molnár et al. 2018). The set of *Gaia* DR2 Cepheids in the SOS results has recently been reclassified (Ripepi et al. 2019), providing a sample with negligible contamination and more accurate Cepheid subclassifications.

4.3. δ Scuti and SX Phoenicis stars

The δ Scuti and SX Phoenicis classifications include 8882 candidates and are merged together (labelled as DSCT_SXPHE) because their *Gaia* light curves can appear very similar

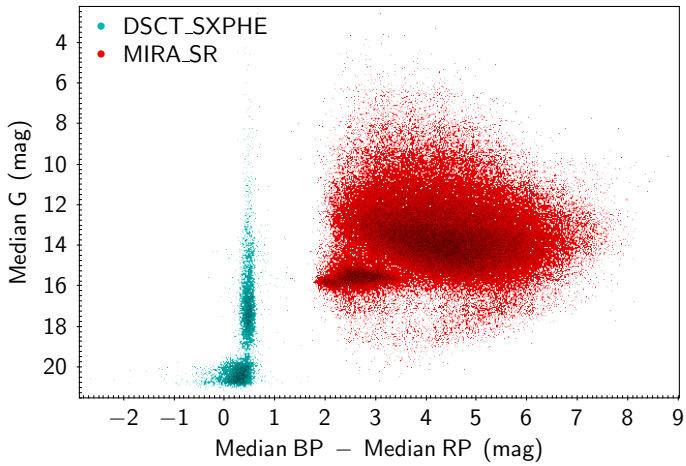


Fig. 25. Colour-magnitude diagram (as median G-band magnitude vs. median $G_{\rm BP}-$ median $G_{\rm RP}$) of the δ Scuti/SX Phoenicis (labelled DSCT_SXPHE, in cyan) and Mira/semiregular (labelled MIRA_SR, in red) classifications.

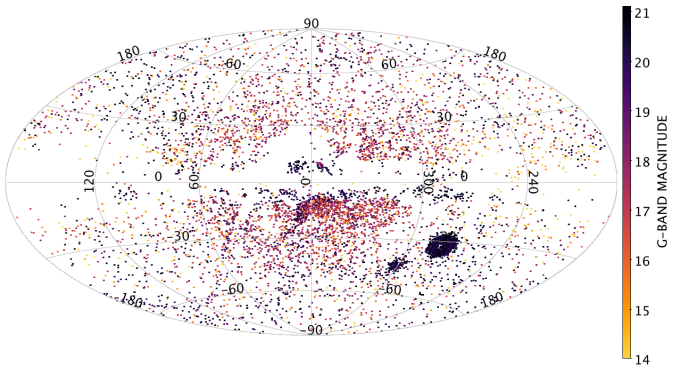


Fig. 26. Same as Fig. 9, but for δ Scuti/SX Phoenicis classifications.

(without metallicity information). Their distribution in a colour-magnitude diagram is shown in Fig. 25 in terms of median G magnitude versus $G_{\rm BP} - G_{\rm RP}$ colour. In general, most of the δ Scuti/SX Phoenicis candidates are identified in the Magellanic Clouds, in the Galactic halo (as expected in particular for the metal-poor SX Phoenicis candidates), and in the Sagittarius dwarf spheroidal galaxy (see Fig. 26). The vast majority of candidates is not affected by Galactic extinction, likely because reddened δ Scuti/SX Phoenicis samples are insufficiently represented in the training set. Approximately half of the candidates are in the clump between median G of 19.5 and 20.7 mag (about the *Gaia* magnitude limit): half of these are located in the Magellanic Clouds, while the others are partly in the Sagittarius dwarf spheroidal galaxy and partly scattered in the Galactic halo. Brighter candidates are distributed mostly in the Galactic halo, with a larger number density in the region below the Galactic bulge, as shown in Fig. 26.

The distribution in the sky of the classification scores of the δ Scuti and SX Phoenicis candidates is presented in Fig. 27, and a comparison with Fig. 26 indicates a definite reduction of score levels in the identifications of candidates fainter than a median G of about 19.5 mag, especially in the Magellanic Clouds and around the Galactic bulge. For sources with median G between 17 and 19.5 mag, the correlation between brighter candidates and higher classification scores is much weaker than at the faint end,

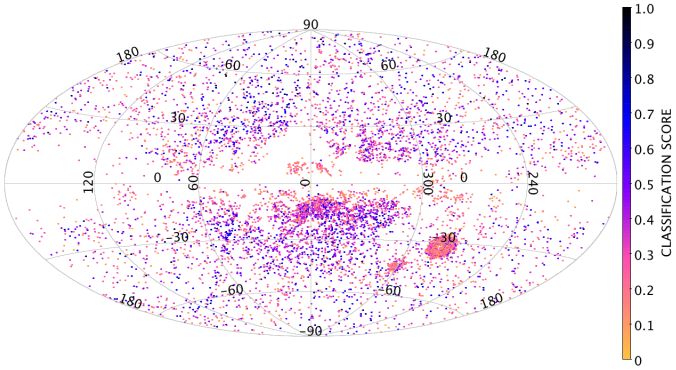


Fig. 27. Same as Fig. 12, but for δ Scuti/SX Phoenicis classifications.

and it becomes negligible for δ Scuti/SX Phoenicis candidates brighter than median $G \approx 17$ mag.

The δ Scuti/SX Phoenicis classifications are not verified by a dedicated SOS module. Although a number of candidates could be confirmed by their period and a comparison of the Fourier amplitudes and phases of the first two harmonics, a large number of insufficiently sampled candidates would remain unconfirmed (but not necessarily rejected). We therefore opted for a simple check of the light variations of the δ Scuti/SX Phoenicis candidates in different bands, and let the community study these candidates in further detail (possibly with additional data). As illustrated for Cepheids and RR Lyrae stars (in Sects. 4.2 and 4.1), δ Scuti/SX Phoenicis pulsators are also expected to exhibit larger variations in the $G_{\rm BP}$ than in the $G_{\rm RP}$ band, although these intrinsically fainter stars have lower amplitudes on average (for a given apparent brightness), and their IQR estimates are therefore more strongly affected by noise. The distribution of the ratios of the IQR in the $G_{\rm BP}$ versus $G_{\rm RP}$ band is shown as a function of median G-band magnitude in Fig. 28. As the IQR is more strongly influenced by the photometric noise especially towards fainter magnitudes, the following fractions of δ Scuti/SX Phoenicis candidates with IQR($G_{\rm BP}$)/IQR($G_{\rm RP}$) > 1 inferred from this distribution represent likely lower limits[6]: 0.87 (=576/664), 0.94 (=1942/2077), 0.91 (=3795/4189), and 0.67 (=5981/8882), for source subsets with median G brighter than 15, 17, 19, and any magnitude, respectively. Higher classification scores further support the same conclusion, as the overall fraction of 0.67 increases to 0.75 (=4449/5944), 0.80 (=3251/4057), 0.86 (=2213/2578), and 0.90 (=1366/1513) in subsets of candidates associated with scores greater than 0.2, 0.3, 0.4, and 0.5, respectively.

As mentioned in Sect. 3.3, all of the cross-matched δ Scuti/SX Phoenicis stars were included in the training set, so that the comparison with the literature employed for Figs. 29–31 is expected to be overly optimistic, but it is still included to show the pitfalls even in the best-case scenario. The apparent completeness and contamination rates of δ Scuti/SX Phoenicis candidates of any classification score are shown in Figs. 29 (in counts) and 30 (in percentage), suggesting that QSOs constitute the main source of contamination. The BLAP classifications were merged with the DSCT_SXPHE class (as explained in

[6] If the number of measurements in $G_{\rm RP}$ is sufficiently smaller than in $G_{\rm BP}$, the former could artificially reduce the IQR($G_{\rm RP}$) and boost the IQR($G_{\rm BP}$)/IQR($G_{\rm RP}$) ratio. However, the distribution of the difference of the number of measurements in $G_{\rm RP}$ and $G_{\rm BP}$ is rather symmetric, relieving concerns of such biases for the fractions of sources with IQR($G_{\rm BP}$)/IQR($G_{\rm RP}$) > 1.

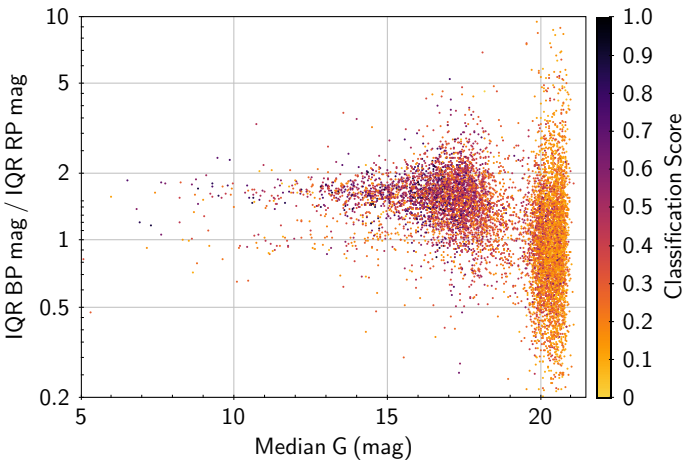


Fig. 28. Distribution of the ratios of the interquartile range (IQR) in the $G_{\rm BP}$ vs. $G_{\rm RP}$ bands as a function of the median G-band magnitude for the δ Scuti/SX Phoenicis candidates. The classification score is colour-coded as indicated in the legend on the right-hand side.

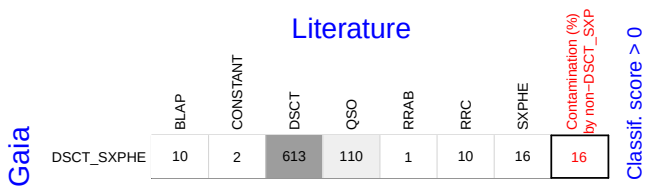


Fig. 29. Same as Fig. 13, but for δ Scuti/SX Phoenicis classifications. In this case, the training-set objects were included in the computation of the completeness and contamination rates.

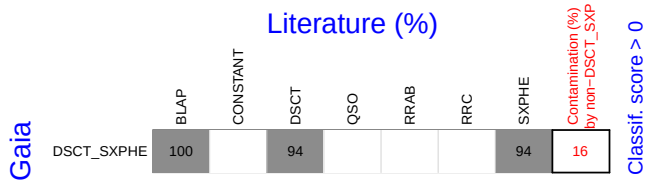


Fig. 30. Same as Fig. 14, but for δ Scuti/SX Phoenicis classifications. In this case, the training-set objects were included in the computation of the completeness and contamination rates.

Sect. 3.2.2), therefore the trained BLAP sources were included among the δ Scuti/SX Phoenicis candidates.

The dependence of the apparent completeness and contamination rates on minimum classification scores for the δ Scuti/SX Phoenicis candidates is shown in Fig. 31. This generally confirms the expected trend that higher score thresholds increase the completeness-to-contamination ratio more efficiently than lower score limits.

4.4. Mira and semiregular variables

The Mira and semiregular classifications include 150 757 candidates of long-period variables (labelled MIRA_SR). Their distribution in a colour-magnitude diagram is shown in Fig. 25 in terms of median G magnitude versus $G_{\rm BP} - G_{\rm RP}$ colour. As apparent in Fig. 32, most of the long-period candidates are associated with the Galactic disc and bulge, which form the largest clump visible in Fig. 25, and approximately 8% of the candidates are in the region of the Magellanic Clouds, which give rise to the secondary clump centred at median $G \approx 15$–16 mag and

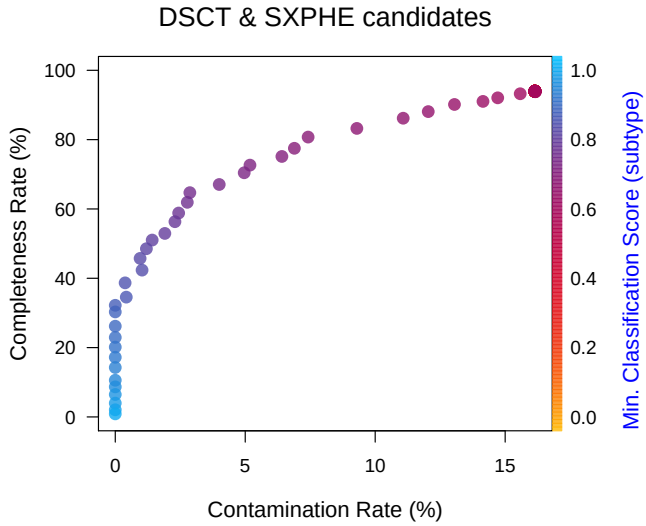


Fig. 31. Same as Fig. 15, but for δ Scuti/SX Phoenicis classifications. In this case, the training-set objects were included in the computation of the completeness and contamination rates.

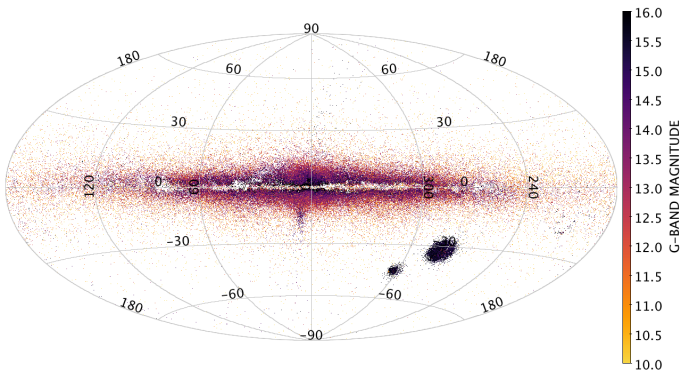


Fig. 32. Same as Fig. 9, but for Mira/semiregular classifications.

bluer than $G_{\rm BP} - G_{\rm RP} \approx 4$ mag. The bright end includes candidates that are distributed more uniformly in the sky, while the faint end is represented primarily by extinguished objects, mostly close to the Galactic equator and some of them in the Magellanic Clouds.

The distribution in the sky of the classification scores of Mira and semiregular variables is presented in Fig. 33 and seems rather uninformative because it is dominated by high score values throughout the sky (about 80% of the candidates have a score greater than 0.8), except for stars close to the Galactic equator (in particular in the region of the bulge), which are classified with reduced confidence.

As mentioned in Sect. 3.3, the selection of reliable Mira and semiregular classifications was particularly challenging because the training set was not fully representative. Instead of filtering contaminating objects out by means of a validation classifier, the Mira and semiregular candidates were passed directly to the SOS module dedicated to long-period variables, and the published subset of these candidates followed from the selections described in Mowlavi et al. (2018). We note that among the contaminants of the Mira and semiregular candidates published in *Gaia* DR2 is a small set (smaller than 1%) of young stellar objects (see details in Mowlavi et al. 2018) and 99 sources that are part of the AllWISEAGN catalogue (Secrest et al. 2015).

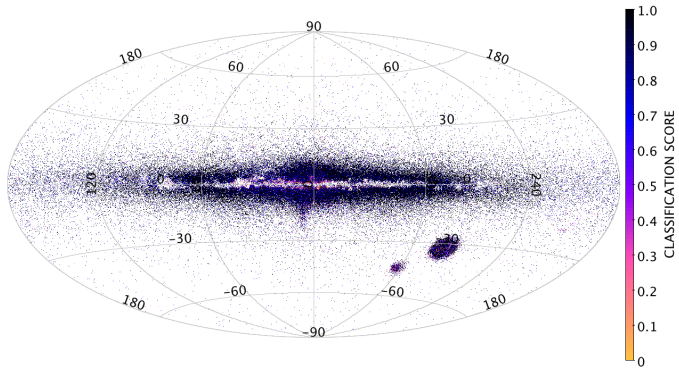


Fig. 33. Same as Fig. 12, but for Mira/semiregular classifications.

5. Conclusions

The first all-sky classifications of four main classes of high-amplitude pulsating stars (RR Lyrae, Cepheid, δ Scuti/SX Phoenicis, and Mira/semiregular) with intermediate *Gaia* data include already large numbers of candidates that can be used in many applications, especially those requiring a combination of a certain degree of homogeneity on the large scales and sufficient statistics of specific variability types. Training-set biases are common in automated supervised classification methods, and their severity was reduced by source sampling (through the large number of sources that could be cross matched with the literature) and semi-supervised techniques. Some of the remaining biases were described for the interpretation of the results of specific (sub)classes, when applicable. The classified subtypes of RR Lyrae and Cepheids can be inaccurate, especially if much less common than others of the same family, without making use of periods, Fourier parameters, or enough observations, but are provided in the attempt to increase the purity of the most common subtype. Parent classes are more reliable, and their use is suggested whenever possible. Validation classifiers helped alleviate the presence of contaminants in the classification results at the cost of some reduction in completeness, and selections of even more reliable candidates were achieved by dedicated SOS modules (Clementini et al. 2019; Mowlavi et al. 2018).

The next *Gaia* data release will include many more variable stars and variability classes, with improved classification accuracy through the increased number of measurements after about three years of observations (making models possible for the majority of sources) and the use of new data types (some of which were already available in *Gaia* DR2, but not early enough to be used in the variability pipeline), in addition to improvements in the astrometric and photometric data. More details on the classification results (such as classification attributes and probability arrays) are also planned to be published in the future.

Acknowledgements. This work has made use of data from the European Space Agency (ESA) mission *Gaia* (https://www.cosmos.esa.int/gaia), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, some of which participate in the *Gaia* Multilateral Agreement, which include, for Switzerland, the Swiss State Secretariat for Education, Research and Innovation through the ESA Prodex program, the "Mesures d'accompagnement", the "Activités Nationales Complémentaires", the Swiss National Science Foundation, and the Early Postdoc.Mobility fellowship; for Belgium, the BELgian federal Science Policy Office (BELSPO) through PRODEX grants; for Italy, Istituto Nazionale di Astrofisica (INAF) and the Agenzia Spaziale Italiana (ASI) through grants I/037/08/0, I/058/10/0, 2014-025-R.0, and 2014-025-R.1.2015 to INAF (PI M.G. Lattanzi); for Hungary, the Lendület grants LP2014-17 and LP2018-7 from the Hungarian Academy of Sciences, and the NKFIH grants K-115709, PD-116175, PD-121203 from the Hungarian National Research, Development, and Innovation Office. L.M. and E.P. have been supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. This work made use of software from Postgres-XL (https://www.postgres-xl.org), Java (https://www.oracle.com/java/), Weka (Frank et al. 2016), R (R Core Team 2018), and TOPCAT/STILTS (Taylor 2005).

References

Aerts, C., & Sterken, C., 2006, Astrophysics of Variable Stars, ASP Conf. Ser., 349
Aerts, C., Christensen-Dalsgaard, J., & Kurtz, D. W. 2010, Asteroseismology (Berlin: Springer)
Andrae, R., Fouesneau, M., Creevey, O., et al. 2018, A&A, 616, A8 (*Gaia* 2 SI)
Arenou, F., Luri, X., Babusiaux, C., et al. 2018, A&A, 616, A17 (*Gaia* 2 SI)
Bellman, R. 1961, Adaptive Control Processes: A Guided Tour (Princeton, New Jersey: Princeton University Press)
Belokurov, V., Erkal, D., Deason, A. J., et al. 2017, MNRAS, 466, 4711
Bradley, P. A., Guzik, J. A., Miles, L. F., et al. 2015, AJ, 149, 68
Breiman, L. 2001, Mach. Learn., 45, 5
Busso, G., Cacciari, C., Carrasco, J. M., et al. 2018, Gaia DR2 documentation Chapter 5: Photometry, Tech. rep
Butler, N. R., & Bloom, J. S. 2011, AJ, 141, 93
Calle, M. L., Urrea, V., Boulesteix, A.-L., & Malats, N. 2011, Hum. Hered., 72, 121
Christensen-Dalsgaard, J. 2004, in SOHO 14 Helio- and Asteroseismology: Towards a Golden Future, ed. D. Danesy, ESA SP, 559, 1
Clementini, G. 2014, in Precision Asteroseismology, eds. J. A. Guzik, W. J. Chaplin, G. Handler, & A. Pigulski, IAU Symp., 301, 129
Clementini, G., Ripepi, V., Leccia, S., et al. 2016, A&A, 595, A133 (*Gaia* 2 SI)
Clementini, G., Ripepi, V., Molinaro, R., et al. 2019, A&A, 622, A60 (*Gaia* 2 SI)
Debosscher, J., Blomme, J., Aerts, C., & De Ridder, J. 2011, A&A, 529, A89
Diaz-Uriarte, R., & Alvarez de Andres, S. 2005, ArXiv e-prints [arXiv:q-bio/0503025]
Drake, A. J., Catelan, M., Djorgovski, S. G., et al. 2013a, ApJ, 765, 154
Drake, A. J., Catelan, M., Djorgovski, S. G., et al. 2013b, ApJ, 763, 32
Drake, A. J., Graham, M. J., Djorgovski, S. G., et al. 2014, ApJS, 213, 9
Evans, D. W., Riello, M., De Angeli, F., et al. 2018, A&A, 616, A4
Eyer, L., Mowlavi, N., Evans, D. W., et al. 2017, A&A, submitted [arXiv:1702.03295] (*Gaia* 2 SI)
Eyer, L., Guy, L., Distefano, E., et al. 2018, Gaia DR2 documentation Chapter 7: Variability, Tech. rep
Frank, E., Hall, M. A., & Witten, I. H. 2016, The WEKA Workbench. Online Appendix for 'Data Mining: Practical Machine Learning Tools and Techniques', 4th edn. (Burlington: Morgan Kaufmann)
Gaia Collaboration (Prusti, T., et al.) 2016, A&A, 595, A1 (*Gaia* 2 SI)
Gaia Collaboration (Clementini, G., et al.) 2017, A&A, 605, A79 (*Gaia* 2 SI)
Gaia Collaboration (Brown, A. G. A., et al.) 2018, A&A, 616, A1 (*Gaia* 2 SI)
Gaia Collaboration (Eyer, L., et al.) 2019, A&A, 623, A110 (*Gaia* 2 SI)
Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. 2010, Pattern Recognit. Lett., 31, 2225
Guyon, I., & Elisseeff, A. 2003, J. Mach. Learn. Res., 3, 1157
Hastie, T., Tibshirani, R., & Friedman, J. 2009, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. (New York: Springer-Verlag)
Heinze, A. N., Tonry, J. L., Denneau, L., et al. 2018, AJ, 156, 241
Hernitschek, N., Schlafly, E. F., Sesar, B., et al. 2016, ApJ, 817, 73
Hernitschek, N., Sesar, B., Rix, H.-W., et al. 2017, ApJ, 850, 96
Holl, B., Audard, M., Nienartowicz, K., et al. 2018, A&A, 618, A30 (*Gaia* 2 SI)
Ivezić, Ž., Goldston, J., Finlator, K., et al. 2000, AJ, 120, 963
Ivezić, Ž., Smith, J. A., Miknaitis, G., et al. 2007, AJ, 134, 973
Jacyszyn-Dobrzeniecka, A. M., & the OGLE Team 2018, Proc. Polis. Astron. Soc., 6, 77
Jayasinghe, T., Kochanek, C. S., Stanek, K. Z., et al. 2018, MNRAS, 477, 3145
Jayasinghe, T., Stanek, K. Z., Kochanek, C. S., et al. 2019, MNRAS, 486, 1907
Katz, D., Sartoretti, P., Cropper, M., et al. 2019, A&A, 622, A205 (*Gaia* 2 SI)
Lanzafame, A. C., Distefano, E., Messina, S., et al. 2018, A&A, 616, A16 (*Gaia* 2 SI)
Lindegren, L., Hernández, J., Bombrun, A., et al. 2018, A&A, 616, A2 (*Gaia* 2 SI)
Long, J. P., El Karoui, N., Rice, J. A., Richards, J. W., & Bloom, J. S. 2012, PASP, 124, 280
Mahalanobis, P. C. 1936, Proc. Nat. Inst. Sci. India, 2, 49
Molnár, L., Plachy, E., Juhász, Á. L., & Rimoldini, L. 2018, A&A, 620, A127 (*Gaia* 2 SI)
Mowlavi, N., Lecoeur-Taïbi, I., Lebzelter, T., et al. 2018, A&A, 618, A58 (*Gaia* 2 SI)
Palaversa, L., Ivezić, Ž., Eyer, L., et al. 2013, AJ, 146, 101

Pawlak, M., Soszyński, I., Udalski, A., et al. 2016, Acta Astron., 66, 421
Pietrukowicz, P., Dziembowski, W. A., Latour, M., et al. 2017, Nat. Astron., 1, 0166
Pojmanski, G. 2002, Acta Astron., 52, 397
R Core Team 2018, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria
Richards, J. W., Starr, D. L., Miller, A. A., et al. 2012, ApJS, 203, 32
Riello, M., De Angeli, F., Evans, D. W., et al. 2018, A&A, 616, A3 (*Gaia* 2 SI)
Rimoldini, L., Nienartowicz, K., Süveges, M., et al. 2017, ArXiv e-prints [arXiv:1702.04165]
Ripepi, V., Molinaro, R., Musella, I., et al. 2019, A&A, 625, A14
Roelens, M., Eyer, L., Mowlavi, N., et al. 2018, A&A, 620, A197 (*Gaia* 2 SI)
ROTSE Collaboration (Kinemuchi, K., et al.) 2006, AJ, 132, 1202
Sandage, A., & Tammann, G. A. 2006, ARA&A, 44, 93
Sartoretti, P., Katz, D., Cropper, M., et al. 2018, A&A, 616, A6 (*Gaia* 2 SI)
Secrest, N. J., Dudik, R. P., Dorland, B. N., et al. 2015, ApJS, 221, 12
Sesar, B., Ivezić, Ž., Lupton, R. H., et al. 2007, AJ, 134, 2236
Sesar, B., Hernitschek, N., Mitrović, S., et al. 2017, AJ, 153, 204
Soszyński, I., Udalski, A., Poleski, R., et al. 2012, Acta Astron., 62, 219
Soszyński, I., Udalski, A., Szymański, M. K., et al. 2014, Acta Astron., 64, 177
Soszyński, I., Udalski, A., Szymański, M. K., et al. 2015a, Acta Astron., 65, 233
Soszyński, I., Udalski, A., Szymański, M. K., et al. 2015b, Acta Astron., 65, 297
Soszyński, I., Udalski, A., Szymański, M. K., et al. 2016a, Acta Astron., 66, 131
Soszyński, I., Pawlak, M., Pietrukowicz, P., et al. 2016b, Acta Astron., 66, 405
Soszyński, I., Udalski, A., Szymański, M. K., et al. 2017, Acta Astron., 67, 297
Souchay, J., Andrei, A. H., Barache, C., et al. 2015, A&A, 583, A75
Taylor, M. B. 2005, in Astronomical Data Analysis Software and Systems XIV, eds. P. Shopbell, M. Britton, & R. Ebert, ASP Conf. Ser., 347, 29
Torrealba, G., Catelan, M., Drake, A. J., et al. 2015, MNRAS, 446, 2251
Woźniak, P. R., Williams, S. J., Vestrand, W. T., & Gupta, V. 2004, AJ, 128, 2965

Appendix A: Training classes

The definitions of training-set class labels are listed below, highlighting the class labels published in *Gaia* DR2 with a bold font.

1. **ACEP**: anomalous Cepheids.
2. ACV: α^2 Canum Venaticorum-type stars.
3. ACYG: α Cygni-type stars.
4. **ARRD**: anomalous double-mode RR Lyrae stars.
5. BCEP: β Cephei-type stars.
6. BLAP: blue large amplitude pulsators.
7. **CEP**: classical (δ) Cepheids.
8. CONSTANT: objects whose variations (or absence thereof) are consistent with those of constant sources.
9. CV: cataclysmic variables of unspecified type.
10. **DSCT**: δ Scuti-type stars.
11. ECL: eclipsing binary stars.
12. ELL: rotating ellipsoidal variable stars (in close binary systems).
13. FLARES: magnetically active stars displaying flares.
14. GCAS: γ Cassiopeiae-type stars.
15. GDOR: γ Doradus-type stars.
16. **MIRA**: long-period variable stars of the o (omicron) Ceti type (Mira).
17. OSARG: OGLE small-amplitude red giant variable stars.
18. QSO: optically variable quasi-stellar extragalactic sources.
19. ROT: rotation modulation in solar-like stars due to magnetic activity (spots).
20. **RRAB**: fundamental-mode RR Lyrae stars.
21. **RRC**: first-overtone RR Lyrae stars.
22. **RRD**: double-mode RR Lyrae stars.
23. RS: RS Canum Venaticorum-type stars.
24. SOLARLIKE: stars with solar-like variability induced by magnetic activity (flares, spots, and rotational modulation).
25. SPB: slowly pulsating B-type stars.
26. SXARI: SX Arietis-type stars.
27. **SXPHE**: SX Phoenicis-type stars.
28. **SR**: long-period variable stars of the semiregular type.
29. **T2CEP**: type-II Cepheids.

Appendix B: Sample ADQL queries

Documentation, examples, and support to access the *Gaia* data are available in the *Gaia* archive help web-page[7]. Sample queries related to the classified variable stars, in some cases after SOS processing, are presented in the appendices of Clementini et al. (2019), Holl et al. (2018), and Mowlavi et al. (2018). We here add a few additional examples for the extraction of classification information combined with astrometry, photometry, or SOS results available in other archive tables.

In order to identify the 618 RR Lyrae classifications that were reclassified as Cepheids in SOS and retrieve information related to the classifier results, the number of G-band FoV transits and their median magnitude, the $G_{\rm BP} - G_{\rm RP}$ colour from the medians of each band, the SOS reclassification labels, and the Galactic latitude and longitude in degrees, the ADQL query is as follows.

```
SELECT c.source_id, best_class_name, best_class_score,
       num_selected_g_fov, median_mag_g_fov,
       median_mag_bp-median_mag_rp AS med_bp_rp,
       type_best_classification, l, b
FROM gaiadr2.vari_classifier_result AS c
INNER JOIN gaiadr2.vari_time_series_statistics AS stat
        ON c.source_id = stat.source_id
INNER JOIN gaiadr2.vari_cepheid AS cep
        ON c.source_id = cep.source_id
INNER JOIN gaiadr2.gaia_source AS s
        ON c.source_id = s.source_id
WHERE best_class_name = 'RRAB' OR
      best_class_name = 'RRC' OR best_class_name='RRD'
      OR best_class_name = 'ARRD'
```

To select δ Scuti/SX Phoenicis candidates with parallax and epoch photometry that satisfy the conditions of IQR($G_{\rm BP}$)/IQR($G_{\rm RP}$) >1.5, median G brighter than 17 mag, classification score greater than 0.5, and relative parallax precision better than 20%, sorted by decreasing IQR ratio, the corresponding ADQL query is as follows.

```
SELECT s.source_id, median_mag_g_fov, best_class_score,
       iqr_mag_bp/iqr_mag_rp AS bp_rp_iqr_ratio,
       parallax, parallax_error, epoch_photometry_url
FROM gaiadr2.vari_classifier_result AS c
INNER JOIN gaiadr2.vari_time_series_statistics AS stat
        ON c.source_id = stat.source_id
INNER JOIN gaiadr2.gaia_source AS s
        ON c.source_id = s.source_id
WHERE best_class_name = 'DSCT_SXPHE' AND
      iqr_mag_bp/iqr_mag_rp > 1.5 AND
      median_mag_g_fov < 17 AND
      best_class_score > 0.5 AND
      parallax_over_error > 5
ORDER BY bp_rp_iqr_ratio DESC
```

[7] http://gea.esac.esa.int/archive-help/index.html