

TESS Data for Asteroseismology (T'DA) Stellar Variability Classification Pipeline: Set-Up and Application to the *Kepler* Q9 Data

J. AUDENAERT ¹, J. S. KUSZLEWICZ,^{2,3} R. HANDBERG ³, A. TKACHENKO ¹, D. ARMSTRONG ^{4,5}, M. HON ^{6,7},
R. KGOADI,⁸ M. N. LUND ³, K. J. BELL ^{9,10}, L. BUGNET ^{11,12}, D. M. BOWMAN ¹, C. JOHNSTON ^{1,13},
R. A. GARCÍA ¹², D. STELLO,^{7,14,3} L. MOLNÁR ^{15,16,17}, E. PLACHY ^{15,16,17}, D. BUZASI ¹⁸, C. AERTS ^{1,13,19}
AND THE T'DA COLLABORATION,

¹*Institute of Astronomy, KU Leuven, Celestijnenlaan 200D, 3001, Leuven, Belgium*

²*Landessternwarte, Zentrum für Astronomie der Universität Heidelberg, Königstuhl 12, 69117, Heidelberg, Germany*

³*Stellar Astrophysics Centre, Department of Physics and Astronomy, Aarhus University, Ny Munkegade 120, DK-8000 Aarhus C, Denmark*

⁴*Department of Physics, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK*

⁵*Centre for Exoplanets and Habitability, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, UK*

⁶*Institute for Astronomy, University of Hawai'i, 2680 Woodlawn Drive, Honolulu, HI 96822, USA*

⁷*School of Physics, The University of New South Wales, Sydney, NSW 2052, Australia*

⁸*College of Science and Engineering, James Cook University, Townsville, Australia, 4811*

⁹*DIRAC Institute, Department of Astronomy, University of Washington, Seattle, WA-98195, USA*

¹⁰*NSF Astronomy and Astrophysics Postdoctoral Fellow*

¹¹*Flatiron Institute, Simons Foundation, 162 Fifth Ave, New York, NY 10010, USA*

¹²*AIM, CEA, CNRS, Université Paris-Saclay, Université Paris Diderot, Sorbonne Paris Cité, F-91191 Gif-sur-Yvette, France*

¹³*Department of Astrophysics, IMAPP, Radboud University Nijmegen, NL-6500 GL, Nijmegen, the Netherlands*

¹⁴*Sydney Institute for Astronomy (SIfA), School of Physics, University of Sydney, Sydney, NSW 2006, Australia*

¹⁵*Konkoly Observatory, Research Centre for Astronomy and Earth Sciences, Eötvös Loránd Research Network (ELKH), Konkoly Thege Miklós út 15-17, H-1121 Budapest, Hungary*

¹⁶*MTA CSFK Lendület Near-Field Cosmology Research Group, 1121, Budapest, Konkoly Thege Miklós út 15-17, Hungary*

¹⁷*ELTE Eötvös Loránd University, Institute of Physics, 1117, Pázmány Péter sétány 1/A, Budapest, Hungary*

¹⁸*Department of Chemistry and Physics, Florida Gulf Coast University, 10501 FGCU Blvd. S., Fort Myers, FL 33965 USA*

¹⁹*Max Planck Institute for Astronomy, Königstuhl 17, 69117 Heidelberg, Germany*

(Received July 15, 2021; Revised July 15, 2021; Accepted July 15, 2021)

Submitted to AJ

ABSTRACT

The NASA Transiting Exoplanet Survey Satellite (TESS) is observing tens of millions of stars with time spans ranging from ~ 27 days to about 1 year of continuous observations. This vast amount of data contains a wealth of information for variability, exoplanet, and stellar astrophysics studies but requires a number of processing steps before it can be fully utilized. In order to efficiently process all the TESS data and make it available to the wider scientific community, the TESS Data for Asteroseismology working group, as part of the TESS Asteroseismic Science Consortium, has created an automated open-source processing pipeline to produce light curves corrected for systematics from the short- and long-cadence raw photometry data and to classify these according to stellar variability type. We will process all stars down to a TESS magnitude of 15. This paper is the next in a series detailing how the pipeline works. Here, we present our methodology for the automatic variability classification of TESS photometry using an ensemble of supervised learners that are combined into a metaclassifier. We successfully validate our method using a carefully constructed labelled sample of *Kepler* Q9 light curves with a 27.4 days time span mimicking single-sector TESS observations, on which we obtain

an overall accuracy of 94.9%. We demonstrate that our methodology can successfully classify stars outside of our labeled sample by applying it to all $\sim 167\,000$ stars observed in Q9 of the *Kepler* space mission.

Keywords: Asteroseismology, Machine learning, Supervised classification

1. INTRODUCTION

Understanding stellar variability is important for many fields of astrophysics. Asteroseismology and stellar astrophysics in general have been revolutionized with the launch of space missions that delivered (and continue delivering) months- to years-long high precision, high cadence, and high-duty cycle brightness measurements for large numbers of stars. Following the MOST (Walker et al. 2003), WIRE (Buzasi 2004; Bruntt & Buzasi 2006) and CoRoT (Auvergne et al. 2009) space missions that were among the pioneers in the field of “space asteroseismology” (e.g. Aerts et al. 2010, for historical notes), *Kepler* (Borucki et al. 2010) observed around 160,000 stars in 30-minute (long) and 1-minute (short) cadence intervals for up to four years. After the failure of its second reaction wheel, the *Kepler* mission was turned into the *Kepler* Second Light (K2; Howell et al. 2014) mission that observed a large number of stars along the ecliptic plane during 20 further campaigns, each of about 80 days duration. The TESS mission (Ricker et al. 2015) was launched in 2018 and is covering almost the entire sky. With millions of stars observed, it offers many times the number of targets as *Kepler* did, but most will be observed for only a fraction of the duration of that mission. The TESS targets in the Full Frame Images (FFIs) were observed at 30-minute cadence intervals during its first 2 years while a pre-selected list of targets was observed at 2-minute cadence. For the first extended mission the FFI cadence is reduced to 10 minutes, with an additional 20 sec cadence introduced as well. The observing periods in a single cycle range from 27.4 d up to 352 d, depending on the position on the sky.

Coping with the large volume of data obtained by various space-missions, and in particular by the currently operational TESS mission, requires a coordinated effort. To that end, the TESS Data for Asteroseismology (T'DA¹) coordinated activity has been created within the TESS Asteroseismic Consortium (TASC²). The major task of the T'DA unit is to serve the community with optimal processing of TESS data (both short cadence and full frame images) for all stars in the sky down to a TESS magnitude of 15. This includes raw light

curve extraction, correction of the extracted light curves for systematics, and their automated classification into variability classes. Putting it into context, thanks to the observing strategy of the TESS mission and a high-level integration of the raw TESS image data into our pipeline which allows us to handle large amounts of data quickly and efficiently, we will ultimately produce an all-sky variability catalogue containing tens of millions of stars. While being a treasure trove on its own, our variability catalogue also forms a rich legacy for future space- and ground-based missions/surveys. The overall scheme of the T'DA operations is depicted in Fig. 1, and includes the data processing and classification pipeline itself as well as the ways our data products are made available to the community. The steps of the light curves extraction and their optimal corrections for systematic effects are described in detail in Handberg et al. (2021) and Lund et al. (in prep.), respectively.

This paper is the next in a series of the T'DA papers and concerns the automated stellar variability classification. This component within the T'DA pipeline structure is highlighted by the red dashed box in Fig. 1, while the classification scheme itself is depicted in Fig. 2. It comprises two major steps: (i) “top-level classification” that is based solely on the information encoded in the light curves themselves; and (ii) “second-level classification” that involves using extra information, such as Gaia parallaxes, photometric colours, etc. This latter classification step also involves using unsupervised methods for variability classification that help us identify potential misclassifications and to search for new (sub-)groups of variable stars within our predefined general variability classes, and will be the subject of a separate future study. The final result is a variability catalog of the whole sky down to a magnitude of 15, containing all the tens of millions of stars observed by TESS. The creation of this large catalog is only possible thanks to the efforts of the entire T'DA team contributing to the pipeline development.

Automated variability classification based on light curves (and frequency spectra) resulted from large-scale surveys such as the Hipparcos mission³, Optical Grav-

¹ <https://tasoc.dk/tda/>

² <https://tasoc.dk/>

³ <https://www.cosmos.esa.int/web/hipparcos>

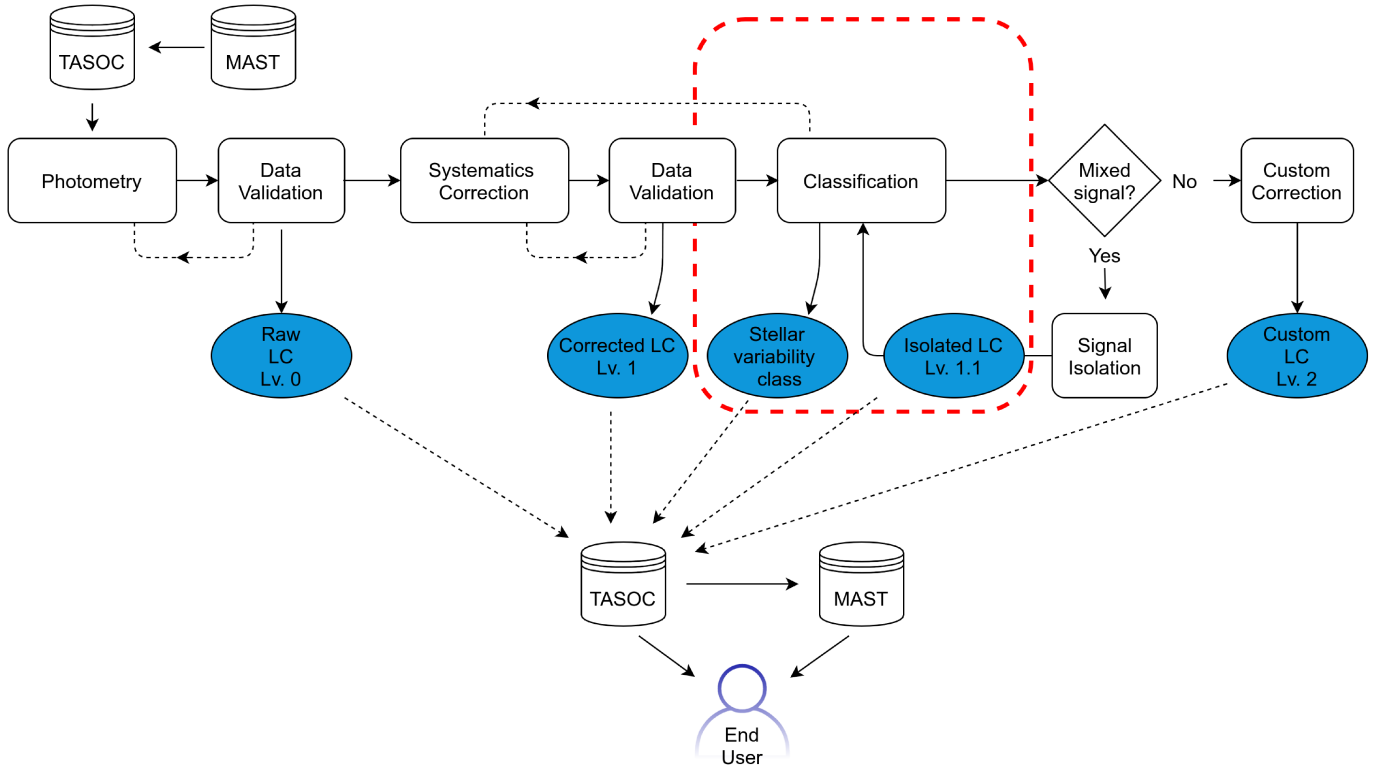


Figure 1. The overall structure of the full T'DA pipeline, with modules given as rectangular boxes, data products as ellipses, and “TASOC” and “MAST” indicate the databases hosting the data products. Dashed lines between modules indicate that an iteration might take place. The part enclosed by the red dashed line indicates the pipeline component described in this paper. The “photometry” part of the pipeline is described in Handberg et al. (2021), while the “correction” is detailed in Lund et al. (in prep.).

itational Lensing Experiment (OGLE⁴), All Sky Automated Survey (ASAS⁵), Sloan Digital Sky Survey (SDSS⁶), etc. The classifications varied in scale from general ones, e.g. Wyrzykowski & Belokurov (2008; OGLE), Pojmanski (2002; ASAS), Ball et al. (2006; SDSS) and Eyer & Grenon (1998; Hipparcos), to those focused on specific types of stars, e.g. Aerts et al. (1998) and Waelkens et al. (1998) from Hipparcos. Debosscher et al. (2007), Sarro et al. (2009), and Debosscher et al. (2011) presented an automated classification of light curves of variable stars in a supervised manner, employing Gaussian Mixtures and Bayesian Networks to classify OGLE, CoRoT, and *Kepler* Quarter 1 (Q1) data. Richards et al. (2011) also used a feature-based approach in combination with a Random Forest to classify variable stars in the OGLE and Hipparcos datasets. More recently, Kim & Bailer-Jones (2016) and Armstrong et al. (2016) respectively used a Random Forest, and Self-Organizing Maps (SOM) in combination with

a Random Forest, to perform classification of variable stars in the ASAS, MACHO (MASSIVE Compact Halo Objects), LINEAR (Lincoln Near-Earth Asteroid Research), and K2 (Camptains 0-4) surveys. Naul et al. (2018) took a hybrid approach and reverted to automated feature learning by means of an unsupervised autoencoder in order to capture the stellar variability, and then subsequently used the latent layer as input into a Random Forest. Jamal & Bloom (2020) extended this approach by making a comprehensive analysis of neural architectures suited for light curve classification. Unsupervised light curve classification is much less prevalent in the literature with a few application examples being by Eyer & Blake (2005), Valenzuela & Pichara (2018) and Modak et al. (2018). Other notable large-scale variability studies include the work by Gaia Collaboration et al. (2016, 2019) for the Gaia mission.

Here, we present a method for the supervised classification of light curves into broad variability classes as depicted by the blue boxes in Fig. 2 (“top-level classification”). We discuss the feature engineering and the collection of the training set, including a detailed description of each variability class, in Sections 2 and 3, respectively. Our individual classifiers are described in

⁴ <http://ogle.astrouw.edu.pl/>

⁵ <http://www.astrouw.edu.pl/asas/>

⁶ <https://www.sdss.org>

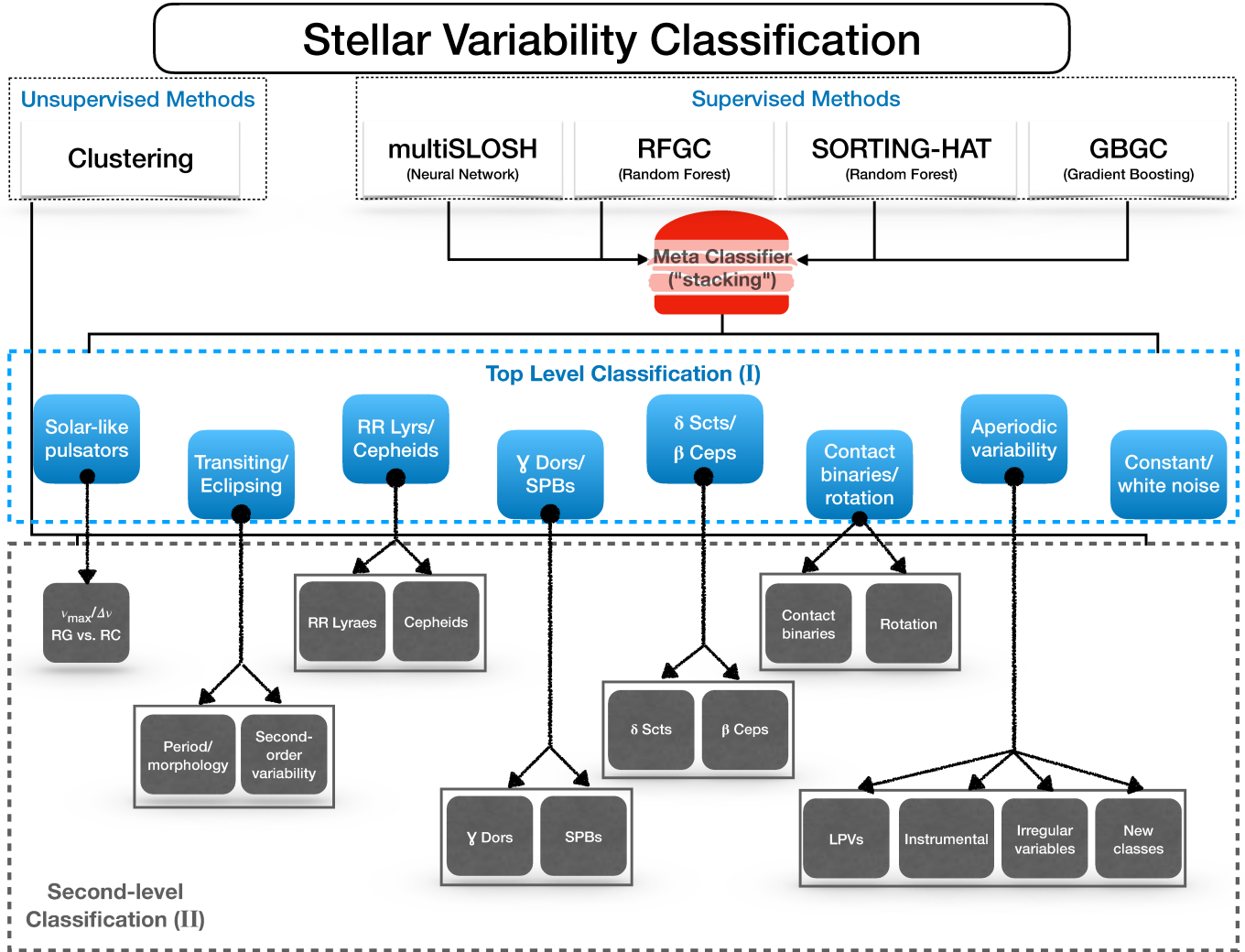


Figure 2. Graphical representation of the TASC classification scheme that encompasses two major stages: “Level 1” is the most general, largely light curve based classification, while “Level 2” stands for a detailed classification based on external features, such as parallaxes, colours, effective temperatures, etc. Rather than only relying on supervised learning, we also make use of unsupervised learning algorithms in Level 2.

Section 4 while their testing and validation is presented in Section 5. The individual classifiers form the basis for the metaclassifier that is tested and validated in Section 6 and is ultimately applied to the truncated 27.4-d segment *Kepler* Q9 data to mimic the single sector TESS case (Section 7). We close the paper with the discussion, conclusions, and an outline of future prospects in Section 8.

2. CLASSIFICATION FEATURES

The optimally extracted and corrected light curves are subject to parameterization; a step that is often referred to as *feature engineering*. The T’DA classification pipeline provides the means for an automated feature extraction that is tuned to the needs of the individual classification algorithms (cf. Sect. 4). Two main types

of features are extracted and used in the process: (i) Fourier-based features and (ii) time-domain features.

2.1. Fourier-based features

An efficient way of extracting periodic signals from a time series of data is to take its Fourier transform. We employ the Lomb-Scargle periodogram method (Lomb 1976; Scargle 1982) to represent input light curves in the Fourier domain and perform classical iterative prewhitening (see e.g., Roberts et al. 1987; Brown et al. 1991; Kjeldsen et al. 1995; Montgomery & Odonoghue 1999; Degroote et al. 2009; Antoci et al. 2019) to extract individual frequencies with their corresponding amplitudes and phases. In this process, stellar flux is repre-

Table 1. Overview of classification features employed by the individual algorithms.

Algorithm/Feature	SLOSH	RFGC	SORTING-HAT	GBGC	Notes
PDS	x				Power density spectrum.
$f_i, j f_i^{(a)}$		x	x	x	Frequencies and their harmonics.
A_{ij}				x	Amplitudes.
$\frac{A_{21}}{A_{11}}, \frac{A_{31}}{A_{11}}$		x			Amplitude ratios.
ϕ_{ij}				x	Phases.
$\phi_{i1} - \phi_{11}, i = 2, 3$		x			Phase differences.
FliPer (F_p) ^(b)					Mean power in a given frequency range
$F_{p0.7, 7, 20, 50}$		x			0.7, 7, 20, 50 μ Hz onwards.
SOM_loc		x			Location on the trained self-organizing maps.
ϕ_{p2p_98}		x			Point-to-point difference, 98 th percentile,
p2p_98		x			ϕ refers to the phase-folded light curve.
ϕ_{p2p_mean}		x			Mean of the point-to-point difference,
p2p_mean		x			ϕ refers to the phase-folded light curve.
ϕ_{range}		x			Range of phase-folded light curve.
D_k		x			Number of zero-crossings in a light curve.
ψ^2		x			Coherency parameter.
$\eta_e^{(d)}$				x	Variability index.
skewness ^(c)			x	x	Light curve skewness.
MAD ^(e)		x		x	Median absolute deviation.
Rcs ^(f)				x	Range of the cumulative sum of the fluxes.
σ^2				x	Variance.
SW ^(g)				x	Shapiro-Wilk test for normality.
kurt ^(h)				x	Kurtosis.
varrat ⁽ⁱ⁾			x		Variance ratio.
SH			x		Number of significant harmonics of f_1 .
FR			x		Flux ratio.
$h(x)$			x		Differential entropy.
MSE					Multiscale entropy
MSE avg,std,max,pow			x		mean, standard deviation, max and power.

^(a) $i \in [1, 6]$ and $j \in [1, 10]$; the number of frequencies and harmonics used is algorithm-dependent

^(b) $F_{p,f_i} = \text{PDS}[f \rightarrow f_{\max}] - P_n$, where P_n is the photon noise computed by considering the averaged power at high frequencies (Bugnet et al. 2018).

^(c) skewness is defined by $\text{skew} = \frac{m_3}{m_2^{3/2}}$, where $m_r = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^r$ is the r_{th} moment about the mean $\bar{x}$

^(d) variability index η_e is computed as ratio of the mean square of successive differences to the variance of the data points

^(e) $\text{MAD} = \text{median}(|X_{0,i} - \text{median}(X_0)|)$, where X_0 stands for the whole time series while the subscript i refers to a single data point in the time series X_0

^(f) cumulative sum of the fluxes is defined by $S_i = S_{i-1} + (x_i - \bar{x})$, $i \in [1, N]$, where $\bar{x}$ is the mean flux

^(g) $\text{SW} = \frac{(\sum_{i=1}^n a_i x_{(i)})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$, where $x_{(i)}$ are the ordered fluxes, $\bar{x}$ the mean flux and a_i the generated constants (see Shapiro & Wilk (1965) for a detailed description)

^(h) kurtosis is defined by $\text{kurt} = \frac{m_4}{m_2^2} - 3$, where $m_r = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^r$ is the r_{th} moment about the mean $\bar{x}$

⁽ⁱ⁾ $\text{varrat} = (\sigma_{init}^2 - \sigma_{sines}^2) / \sigma_{init}^2$, where $\sigma_{sines}^2 = \sum_{i=1}^j A_i^2$, A the amplitude and j the number of harmonics

sented as

$$X(t) = C + \sum_{i=1}^n \sum_{j=1}^m (a_{ij} \sin(2\pi f_i j t) + b_{ij} \cos(2\pi f_i j t)), \quad (1)$$

with C representing the mean value of flux, n and m are the number of extracted frequencies f_i and their harmonic terms, and a_{ij} and b_{ij} are the Fourier coefficients. The coefficients are converted into time-translation invariant frequency amplitude (A_{ij}) and phases (ϕ_{ij}) (see e.g., Bracewell 1986; Aerts et al. 2010).

Iterative prewhitening assumes obtaining and subtracting the optimal fit for the frequency f_i and its m harmonic terms from the flux $X(t)$, and repeating the procedure until n frequencies are extracted from the data. The total number of extracted frequencies varies between individual time series and is determined by a significance criterion. This criterion can be based on an amplitude signal-to-noise or on a threshold calculated in the Fourier domain, as in e.g. Pápics et al. (2012). Such an approach prevents the extraction of spurious frequencies which are the residual signal from the preceding prewhitening steps. We refer the reader to Van Reeth et al. (2015a) and Antoci et al. (2019) for a detailed discussion on the method. The obtained set of frequencies, amplitudes, and phases form a basis for calculation of the Fourier-based classification features whose overview is provided in Table 1. In Fig. 3 we show as an example the ability of Fourier attributes f_1 and f_2 to separate the different classes. It is clear from this that the dSct/bCep class is the most well separated, followed by the gDor/SPB class. The latter does have a small but non-negligible overlap with contactEB/spots stars. In general we see a good structure in the distribution, but it is far from perfect. In order to obtain good classifications they are therefore complemented with other Fourier and non-Fourier attributes.

The Fourier-based feature selection assumes periodic signals as a good representation of the light curve. This is however not particularly suitable for stars that exhibit either stochastic variability or no variability within the detection limit of an instrument. For that reason, some of our individual algorithms work with image-like features where the power density spectrum (PDS) is represented as an image. The PDS is the dominant frequency analysis method for stochastically excited oscillations (Hekker & Christensen-Dalsgaard 2017; García & Ballot 2019). The multiclass solar-like oscillation shape hunter algorithm (multiSLOSH, see Sect. 4.1 for details) therefore performs image recognition on the PDS of variable stars.

2.2. Time-domain features

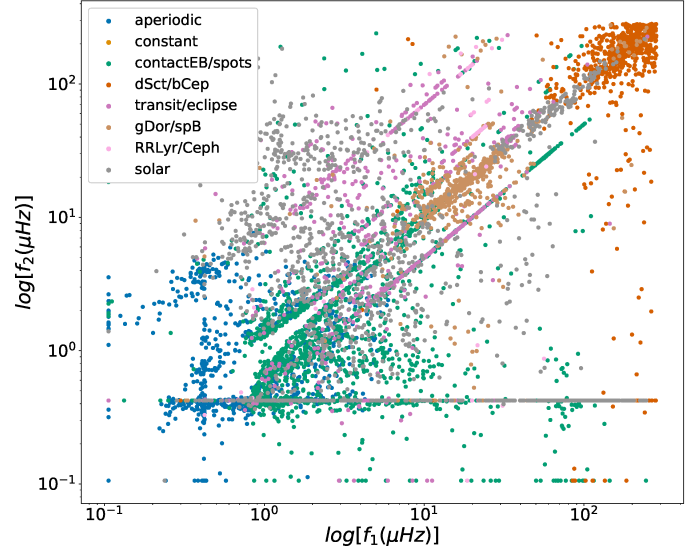


Figure 3. Scatter plot of the full training set for $\log(f_1)$ and $\log(f_2)$ colored per variability class in our classification scheme as defined in Table 2.

Other classification features are extracted directly from the time series and are statistical measures of the distribution of data points in the time series. Some of those are well-known, general statistics features (e.g., skewness and variance). Below we provide a short description of features that are less intuitive and hence require a certain level of insight. All time-domain features are listed in Table 1 with the reference to classifiers that use them.

The *zero-crossings* parameter is computed from the “clipped” time series $Z_{k,i}$, defined as

$$Z_{k,i} = \begin{cases} 1 & \text{if } X_{k,i} \geq 0 \\ 0 & \text{if } X_{k,i} < 0, \end{cases} \quad (2)$$

where $X_{k,i}$ stands for the input time series comprising N data points and with a mean of zero, and k for the k^{th} order difference (see next paragraph). The number of zero-crossings D_k is then computed directly from the “clipped” time-series and is given by

$$D_k = \sum_{i=2}^N (Z_{k,i} - Z_{k,i-1})^2. \quad (3)$$

We normalize the number of zero-crossings to the total number of points N in the light curve to account for a possibly different length of the time series for the individual targets. Setting $k = 0$ gives the number of zero-crossings in the original light curve while $k > 0$ refers to the number of zero-crossings in the time series of higher-order differences. The k^{th} order differences is

defined by

$$X_{k,i} = X_{k-1,i} - X_{k-1,i-1}. \quad (4)$$

For example, the 1st order differences $X_{1,i}$ is given by the point-to-point differences in the original time series X_0 , the 2nd order differences $X_{2,i}$ is given by the point-to-point differences in the time series X_1 , and so on (Kuszelewicz et al. 2020; Kedem & Slud 1981; Kedem & Slud 1982; Bae et al. 1996).

The *coherency* parameter ψ^2 is a measure of the coherence (or stochasticity) of the signal in a time series and is computed from the zero-crossings of the higher-order differences in the time series. It is given by

$$\psi^2 = \sum_{k=0}^5 \frac{(\Delta_k - \phi_k)^2}{\phi_k}, \quad (5)$$

where Δ_k gives the rate of change (i.e. increments) of the number of zero-crossings in the time series of higher-order differences and $\phi_k = (0.167, 0.066, 0.038, 0.025, 0.018)$ are the increments computed from simulated time series of white noise (Kuszelewicz et al. 2020).

The *flux ratio* is the ratio of the sum of squared residuals of the fluxes either brighter or fainter than the mean flux (Kim & Bailer-Jones 2016) and is meant to capture eclipse-like variability. It is defined as

$$\text{FR} = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2}{\frac{1}{M} \sum_{j=1}^M (x_j - \bar{x})^2}, \quad (6)$$

where $\bar{x}$ is the mean flux of the light curve and x_i and x_j the fluxes respectively brighter or fainter than the mean flux. For sinusoidal light curves the ratio is close to unity, while for light curves with eclipses the steep flux gradients cause it to be larger than unity.

The *differential entropy* is an extension of the Shannon Entropy (Shannon 1948) into the continuous domain. It is a measure of the average uncertainty of a variable, and thus a quantification of its unpredictability. The Shannon entropy $H(x)$ of a discrete random variable x is defined as

$$H(x) = - \sum_{i=1}^n p(x_i) \log p(x_i) = -E[\log p(x_i)], \quad (7)$$

where E is the expected value.

We use the differential entropy because, although the light curves are not continuous, they can typically take on a large range of values, causing the number of discrete states to equal the number of samples. This could distort the calculation in the discrete case, so we therefore opted to use the differential entropy. As an alternative we could have opted to use a binned version of

the Shannon entropy. The differential entropy $h(x)$ of a continuous random variable x is defined as

$$h(x) = - \int \mu(x) \log(\mu(x)) dx, \quad (8)$$

where $\mu(x)$ is the density function.

The entropy $h(x)$ can be calculated for a light curve or power density spectrum, where in the latter case it essentially becomes the spectral entropy. Although both are strongly correlated, they complement each other in specific areas. The calculations of $h(x)$ are done with the Python-based Non-parametric Entropy Estimation Toolbox (NPEET)⁷, which uses the Kozachenko-Leonenko estimate (Kozachenko & Leonenko 1987) to calculate the differential entropy as defined in Kraskov et al. (2004).

The sample entropy (Richman & Moorman 2000) is a different type of entropy metric that evaluates the complexity of a time series. The Sample entropy S_E of a signal is defined as

$$S_E(m, N, r) = -\ln \frac{A}{B} = \ln \frac{\sum_{i=1}^{N-m} n_i^m}{\sum_{i=1}^{N-m} n_i^{m+1}} \quad (9)$$

where m is the number of consecutive data points or the embedding dimension, r the tolerance, N the number of data points and n_i the number of vectors close to a basis vector, i.e. $d[u_i^m, u_i^{m+1}] \leq r$.

In practice we calculate the sample entropy by first identifying all unique sequences consisting of m consecutive data points, where each data point is written as $x_i + r$, with r a tolerance margin usually set to a factor of 0.15 of the time series standard deviation. We then count how many times a sequence or template vector of length m occurs and subsequently extend the template vector to length $m + 1$ and count how many times that occurs. The calculations are repeated for each of the next m and $m + 1$ template vector to determine the ratio between the total number of m and $m + 1$ component templates, A and B respectively in Eq. (9). The sample entropy is the natural logarithm of this ratio and represents the probability that a sequence matching each other for the first m data points also match for the next $m + 1$ data points.

The *Multiscale entropy* (MSE; Costa et al. 2005) takes advantage of the fact that stellar variability is active on multiple time scales. Rather than calculating one entropy metric for the full series, we calculate the entropy at each time scale, allowing us to capture the full complexity. More specifically, we first coarse-grain the

⁷ <https://github.com/gregversteeg/NPEET>

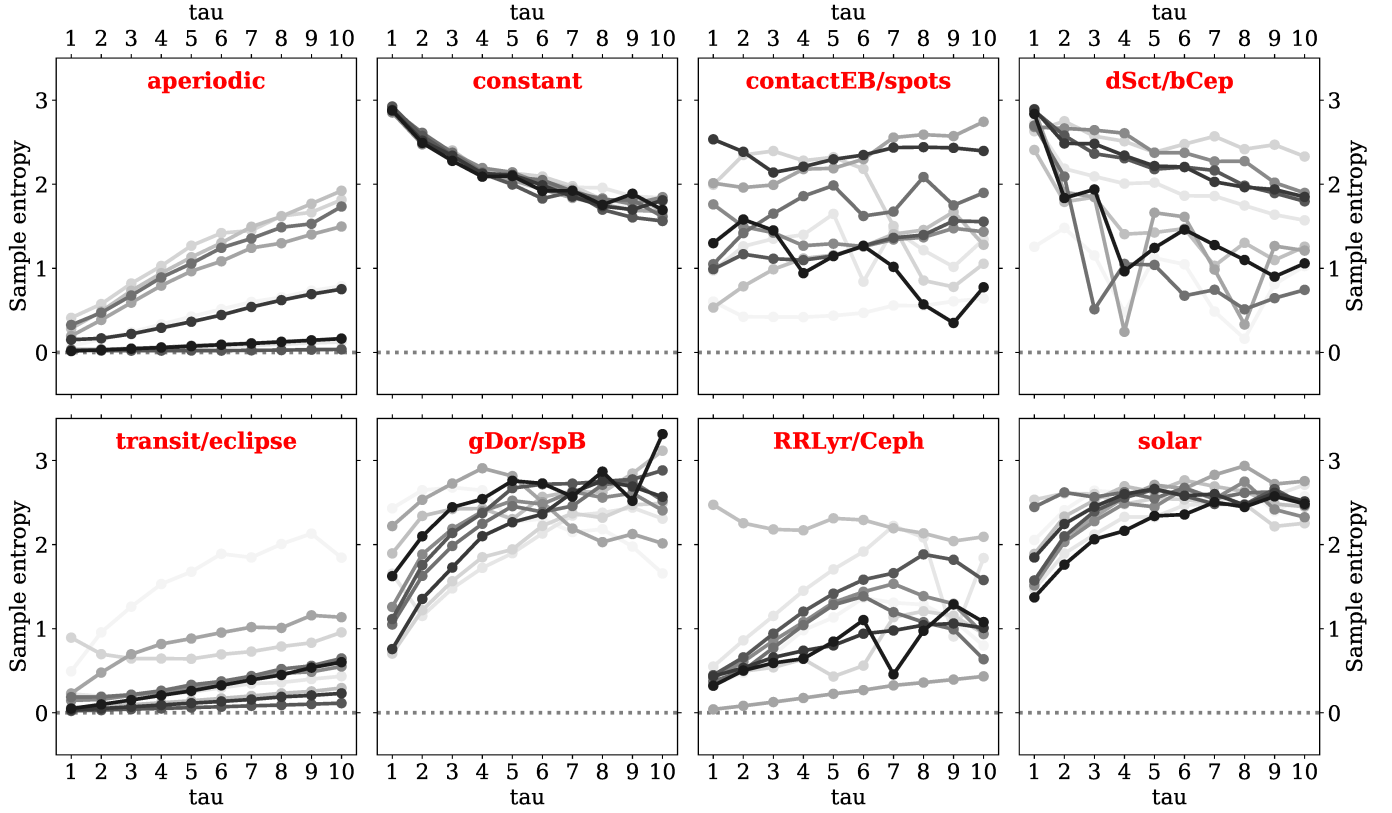


Figure 4. Examples of the Multiscale Entropy (MSE) curves for 10 random samples per variability class in our classification scheme as defined in Table 2.

signal and then calculate the sample entropy for each of these new signals. This allows the MSE to assign minimum values to both deterministic/predictable signals and random/unpredictable signals. Given a time series $x_1, \dots, x_i, \dots, x_N$, the coarse-graining is achieved by dividing the time-series into non-overlapping windows of length τ . Each element x_j in this new time series is then calculated as

$$x_j^\tau = \frac{1}{\tau} \sum_{i=(j-1)\tau+1}^{j\tau} x_i, \quad 1 \leq j \leq \frac{N}{\tau}, \quad (10)$$

where τ is the window length, N the time series length and j the index after coarse-graining.

For $\tau = 1$, the time series $\{x_j^\tau\}$ is simply the original series. For each coarse-grained time series we then calculate the sample entropy given by Eq. (9) and plot it as a factor of the scale. The different types of complexity will then be represented by different types of MSE curves. In general we can say that 1) if for most values of τ the entropy is higher for one signal than for another, that signal is considered more complex, and 2) that a monotonic decrease of the entropy curve indicates that the signal only contains information on the shortest time scale. This monotonic decrease is exactly what we notice in the case of uncorrelated random signals (i.e.

white noise in constant stars), as they only contain information on the shortest time scale, while in other signals information is often present across multiple time scales.

In order to obtain consistent Sample Entropy values it is suggested to have 200 data points per window at the minimum (Busa & van Emmerik 2016). Given that the shortest light curves observed by the TESS nominal mission will have a time span of ~ 27.4 days, consisting of slightly over 1300 data points, we set $\tau_{max} = 10$. This means that for the majority of the coarse-grained time series we have more than 200 data points, where at the smallest window length, i.e. when the scaling factor reaches 10, we have around 130 data points, which is still acceptable in terms of stability. We also did experiments with $\tau_{max} = 20$, and those provided good results as well. Fig. 4 shows the MSE curves for ten random samples per variability class. The figure illustrates the MSE's separating capacity, in particular for constant stars, solar-like oscillators and gDor/SPB stars. Due to complexity associated with implementation of the full curves, we parametrize MSE through its max-

Table 2. Description of the training set.

Class label	Type	Size
aperiodic (Sect. 3.1)	Aperiodic stars	830
contactEB/spots (Sect. 3.2)	Contact binaries and rotational variables	2 260
dSct/bCep (Sect. 3.3)	δ Sct and β Cep stars	772
transit/eclipse (Sect. 3.4)	Eclipsing binaries	974
gDor/SPB (Sect. 3.5)	γ Doradus and SPB stars	630
RRLyrr/Ceph (Sect. 3.6)	RR Lyraes and Cepheids	62
solar (Sect. 3.7)	Solar-like pulsators	1 800
constant (Sect. 3.8)	Constant stars	1 000

imum, mean, standard deviation and power⁸, and use these as classification features.

Lastly, the random forest general classification algorithm (RFGC, see Sect. 4.2 for details) employs the location of a star on the self-organising map (SOM; Kohonen 1990) as one of the features in its classification scheme. The SOM location is obtained by comparing light curve shapes after folding them on the dominant extracted period, essentially grouping similar shapes into clusters.

3. VARIABILITY CLASSES & TRAINING SET

The scientific needs of the TESS Asteroseismic Consortium drive the selection of the main variability classes (schematically represented in Fig. 2) and hence our selection of the training set. Below we provide a short description of each of the variability classes listed in Table 2 alongside the selection criteria that were used to select stars into the respective classes. We made sure to, where possible, maintain a balanced distribution across the different classes, while still incorporating more stars for those classes for which larger known samples exist. For all but one (constant, see below for details) variability classes, we make use of the latest Kepler data release 25⁹ (Thompson et al. 2016), specifically the first 27.4 days of the Q9 PDCSAP data. Our choice of the 27.4 days total time base is dictated by the length of the majority of TESS data – two full orbits of the satellite around the Earth. The choice of the total length of the light curve and building the training set from white-light space-based *Kepler* photometric data enables a

smooth knowledge and methodology transition to the TESS data afterwards. The choice for Q9 was made because it has the least gaps of all Kepler quarters.

3.1. Aperiodic variables

Aperiodic variability (aperiodic) is a class introduced to account for targets whose variability (for one reason or another) appears to be lacking periodicity over time scales shorter than 27.4 days. For example, these can be Mira long-period variables whose variability remains unresolved on the time scale of 27.4 days as only a small fraction of the variability cycle is being captured. Similarly, a fraction of rotational variables may also appear as aperiodic stars due to their rotation periods being much longer than the length of the data set.

Our selection of aperiodic variables is based on the catalog of long-period variables compiled by Yu et al. (2020). The selection consists of 830 objects with *Kepler* Q9 data and having periods longer than 13.7 days so that less than two variability cycles are covered on the time scale of 27.4 days. An example of the light curve and amplitude spectrum of a *Kepler* aperiodic variable is shown in Fig. 5 (first row).

3.2. Contact binaries & rotational variables

Contact binaries and rotational variables (contactEB/spots) is a combined class of *i*) contact binary systems, and *ii*) objects whose light curves show signatures characteristic of surface inhomogeneities modulated by stellar rotation over time. Contact binaries are short-period gravitationally bound systems of two stars that both fill their Roche-lobes, and are therefore in contact at the Lagrangian point L₁. An example of a rotational variable is that of chemically peculiar B, A, F-spectral type stars that show anomalies in their surface chemical composition often associated with a non-uniform distribution of chemical elements. These surface inhomogeneities of either enhanced or depleted abundances of certain chemical elements are often termed “spots” as they appear to a distant observer as darker/brighter regions with respect to the bulk of the star due to significantly modified local opacities (Preston 1974).

The term “surface spots” in application to B, A, F-spectral type stars with radiative envelopes should not be confused with surface spots observed in cooler stars that have extended convective envelopes, e.g. in the Sun. In the latter case, these are regions of reduced surface temperature associated with the contribution of a magnetic field to the total pressure, reducing the gas pressure. Solar-type spots are typically short-lived and vary in their appearance on the time-scales ranging from

⁸ $\text{MSE}_{\text{power}} = \frac{1}{\tau} \sum_{i=1}^{\tau} S_E^2$

⁹ <https://archive.stsci.edu/missions-and-data/kepler/documents/data-release-notes>

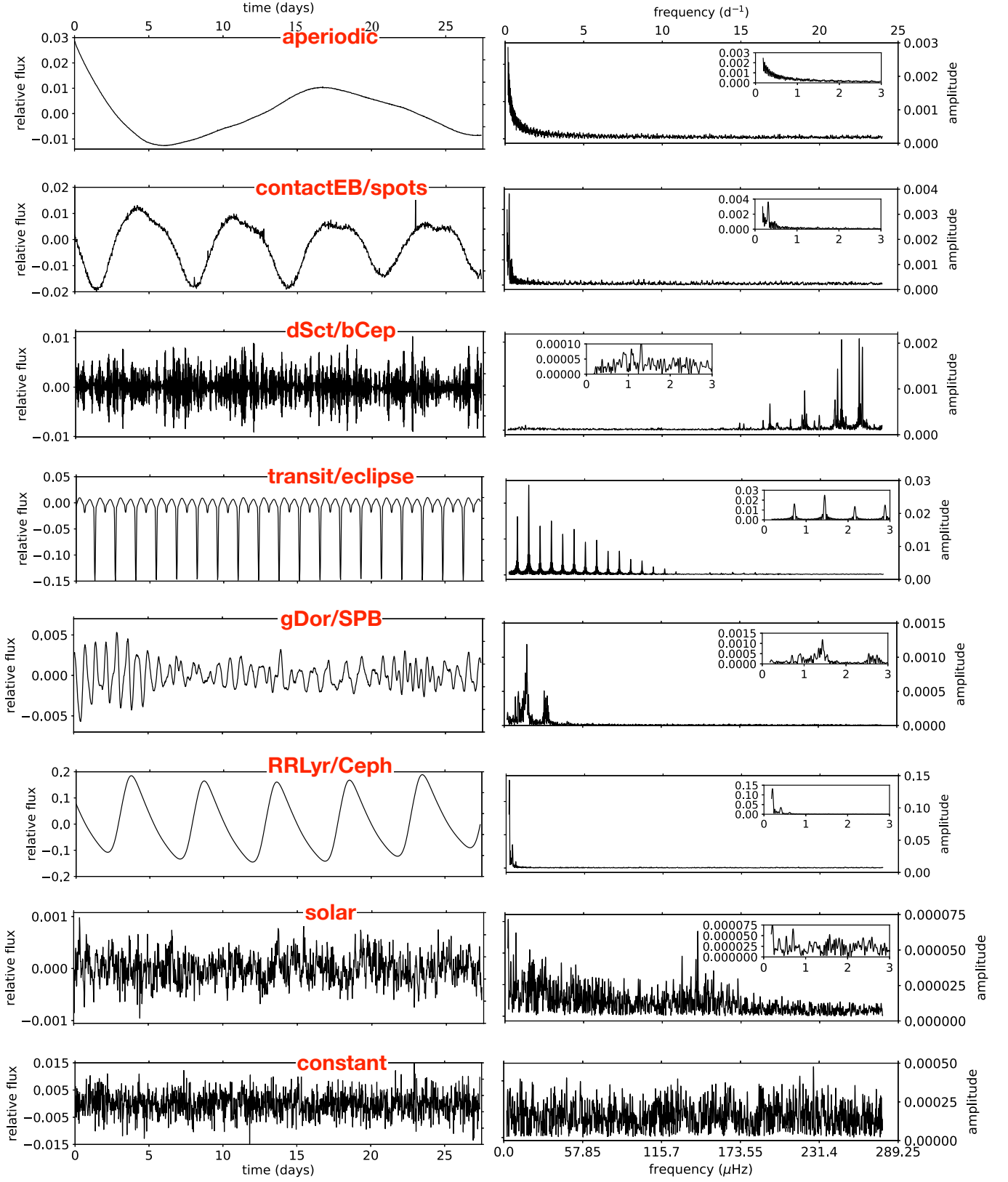


Figure 5. Examples of the light curves (left column) and the respective amplitude spectra (right column) from the training set as defined in Table 2. The inset in the amplitude spectrum panel (where provided) shows a zoom-in into the low-frequency domain of 0 to 3 d^{-1} . Note the different scale on the Y-axis.

a day to a few months (McQuillan et al. 2014; García et al. 2014; Santos et al. 2019). Depending on the level of stellar magnetic activity, such “temperature spots” can be covering up to a few percent of the stellar surface, hence notably modulating the light curves of the respective stars (e.g. Namekata et al. 2019). Because many spots with varying temperature gradients and surface areas can be formed at the same time, light curves of cool active stars are typically much more complex than those of B, A, F-spectral type chemically peculiar stars whose “chemical abundance spots” are long-lived (i.e. timescales ranging from years to decades; Mathys et al. 2020).

Our selection of rotational variables of cool stars is based on the catalog by McQuillan et al. (2014). The catalog contains rotation periods measured for over 30 000 *Kepler* main-sequence stars and selected to have $KIC(T_{\text{eff}}) < 6500$ K. In order to make sure at least two rotation cycles are covered with the 27.4 days data, we restricted our selection to systems whose rotation periods are shorter than 13.7 days. A total of 907 objects all having *Kepler* Q9 light curves were selected this way. The training set was enriched with rotational variables of hotter stars, i.e. with stars having $KIC(T_{\text{eff}}) \geq 6500$ K. For this, we used the catalogs by Nielsen et al. (2013) and Hümmerich et al. (2018) which were cross-matched with the lists of dSct/bCep and gDor/SPB variables (see below) to check for and remove possible duplicates. Furthermore, we excluded stars that do not have *Kepler* Q9 data and/or whose rotational modulation signal is nowhere near the dominant signal in the light curve/amplitude spectrum. A total of 656 objects passed the above selection criteria and were added to the list of 907 cool rotational variables. An example of the light curve and amplitude spectrum of a rotational variable is shown in Fig. 5 (second row).

By analogy with the transit/eclipse class (see below), we queried the *Kepler* Eclipsing Binary Catalog for stars that have *Kepler* Q9 data and whose light curve morphology parameter is larger than 0.6 (high probability contact systems according to Matijević et al. 2012), given that their light curve morphologies look similar to rotational variables. All 1054 systems selected that way were subject to visual inspection to remove misclassified stars of semi-detached type, resulting in the final selection of 697 contact binaries. Altogether, the contactEB/spots class comprises 2 260 objects, of which 70% are rotational variables.

3.3. δ Scuti & β Cephei stars

δ Sct and β Cep (dSct/bCep) stars are two classes of variables pulsating in radial and low-order non-radial

pressure (p) and gravity (g) modes which are mostly excited by means of the κ mechanism acting on the zone of partial ionization of helium (δ Sct stars) and of iron-group elements (β Cep stars, Aerts et al. 2010). The instability regions of the δ Sct and β Cep stars (partially) overlap in the HR diagram with those of γ Dor and SPB stars (see below), respectively, giving rise to hybrid pulsators that exhibit both low-order p modes and high-order g modes simultaneously. β Cep stars have masses between 8 and 25 $M_{\odot}$ and the periods of their pulsations range from about 2 to 8 hours, although none were observed by the *Kepler* mission (Bowman 2020). Less massive δ Sct stars cover the mass range from 1.5 to 2.5 $M_{\odot}$ and have periods from some 15 minutes to about 8 hours (Aerts et al. 2010), hence a significant overlap with β Cep stars in terms of pulsation periods.

It is difficult to distinguish between δ Sct and β Cep stars solely based on their light curve information. Hence, we introduce a joint class of coherent p-mode (δ Sct/ β Cep) pulsators in our classification scheme. Our selection of the training set for this class is based on the δ Sct catalog compiled by Bowman et al. (2016). All 983 objects from that catalog were cross-matched with the catalogs we used to select g-mode pulsators (see below) to search for and remove possible duplicates. Light curves of the remainder of stars were subject to a visual inspection in order to exclude objects with pronounced signatures of rotational modulation as well as stars whose dominant pulsation signal was found to be in the g-mode regime (those hybrid pulsators were included in the class of g-mode pulsators; see below). Ultimately, we selected 772 objects into the class of p-mode pulsators, among those are stars showing p-modes only and hybrid pulsators whose dominant signal is in the p-mode frequency domain. An example of the light curve and amplitude spectrum of a *Kepler* δ Sct p-mode pulsator is shown in Fig. 5 (third row).

3.4. Eclipsing binaries and transit events

Eclipsing/Transiting (transit/eclipse) systems are a class of objects that show extrinsic variability in the form of periodic transits/eclipses. The latter occur due to a partial or total obscuration of the stellar disk by the companion that can be of either stellar (eclipses) or a planetary (transit) mass. We do not make a distinction between transits and eclipses, neither do we intend to distinguish between binary/multiple stellar systems with different Roche geometries (e.g., detached or semi-detached configurations, etc.). Instead, we introduce a general class of eclipsing/transiting objects in our classification scheme which is also likely to contain members whose stellar components are intrinsically variable stars.

Many of eclipsing and transiting systems have been discovered in the *Kepler* space-photometry in recent years, with a variety of orbital and stellar/planetary configurations. The most up-to-date overview of the detections in the *Kepler* field can be obtained from the *Kepler* Eclipsing Binary Catalog¹⁰ and from the NASA Exoplanet Archive¹¹.

Our selection of the training set for the transit/eclipse class is based on the latest release of the *Kepler* Eclipsing Binary Catalog (Prša et al. 2011; Slawson et al. 2011; Kirk et al. 2016; Abdul-Masih et al. 2016). We started by selecting all systems with the morphology parameter smaller than 0.6 which allows us to filter out contact binaries while keeping the majority of detached and semi-detached systems (Matijević et al. 2012). The light curves of all those 1679 objects were subject to a visual inspection in order to remove i) systems whose eclipses are hidden in the noise or any other astrophysical signal and are not traceable in the time domain without aggressive cleaning of the light curve; and 2) (long-period) systems that do not show a single eclipse event in the first 27.4 days segment of their *Kepler* Q9 light curve. Our final training set for the class comprises 974 objects; an example of the light curve and amplitude spectrum of a *Kepler* eclipsing binary is shown in Fig. 5 (fourth row).

3.5. γ Doradus & Slowly Pulsating B stars

γ Dor and Slowly Pulsating B (SPB) stars (gDor/SPB) are members of a class of high non-radial order g-mode pulsators whose oscillations are excited by means of the flux blocking mechanism at the base of their convective envelope (γ Dor stars; Guzik et al. 2000) and by means of the κ mechanism operating on the zone of partial ionization of iron-group elements (SPB stars; Aerts et al. 2010). Although γ Dor and SPB stars occupy different locations in the HR diagram representing F- (mass range between some 1.2 and 2.0 $M_{\odot}$) and B- (with masses from some 3 to 9 $M_{\odot}$) type stars, respectively, their light curves are remarkably similar. The light curves of γ Dor and SPB stars are shaped by an ensemble of g-mode pulsations whose periods range from ~ 0.2 to ~ 3 days.

Our selection of g-mode pulsators is based on several intermediate- to large-scale studies of F- and B-type stars in the *Kepler* field. The sample of lower mass γ Dor stars was adopted from Tkachenko et al. (2013); Van Reeth et al. (2015a,b, 2016); Li et al. (2020), making sure to cross-match between the catalogs to exclude

possible duplicates. In addition, the catalog of δ Sct stars compiled by Bowman et al. (2016) was used to complement pure g-mode pulsators with stars that show both g- and p-modes simultaneously, the so-called hybrid pulsators. We selected only those hybrid pulsators from Bowman et al. (2016) whose dominant variability was found in the g-mode frequency domain. Finally, the training set of g-mode pulsators was enlarged with SPB stars from Pápics et al. (2017) and Pedersen et al. (2020), providing us with a total of 694 stars, of which 630 objects have *Kepler* Q9 data. Because of the similar observational properties of their light curves, we do not distinguish γ Dor stars from their higher-mass SPB counterparts and combine them into a joint class of g-mode pulsators in our classification scheme. A typical light curve and amplitude spectrum of a g-mode pulsator is shown in Fig. 5 (fifth row).

3.6. RR Lyrae and Cepheid stars

Classical pulsators (RRLyr/Ceph class) are low- to intermediate-mass evolved stars whose intrinsic pulsation variability is driven by the opacity (κ) mechanism acting on the partial ionisation zone of helium. The majority of these stars pulsate in a single dominant radial mode and have characteristic non-sinusoidal light curves. However, a small fraction of these objects show two or even three radial modes with comparable amplitudes. Variability of RR Lyrae stars occurs at periods shorter than 1 day, while Cepheids cover a much larger period range, from half a day to several months.

About 50 RR Lyrae stars were identified in the *Kepler* field during the mission (Szabó 2018). In Q9, 42 of those were observed: 34 fundamental-mode and 8 first-overtone pulsators. No double-mode RR Lyrae stars have been targeted in the field, and only two Cepheids have been confirmed: a classical Cepheid, V1154 Cyg, and a medium-period, type II Cepheid, DF Cyg (Szabó et al. 2011; Drekas et al. 2017; Kiss & Bódi 2017; Vega et al. 2017; Plachy et al. 2018; Manick et al. 2019). From the list of 44 RR Lyrae stars and Cepheids, we excluded one object whose 27.4 days segment of the *Kepler* light curve and Fourier transform do not display any significant signal. To increase the training sample, we collected 19 further Cepheids from K2 observations, and created artificial light curves for them. We extrapolated the Fourier decomposition of the light curves to the Q9 time stamps and added appropriately scaled white noise to the data. Together, the 19 simulated Cepheid-type light curves and 43 *Kepler* Q9 RR Lyrae/Cepheid light curves provide us with a total of 62 objects in the final training set for the class. We do not differentiate between RR Lyrae stars and Cepheids in our classifica-

¹⁰ <http://keplerebs.villanova.edu>

¹¹ <https://exoplanetarchive.ipac.caltech.edu>

tion scheme, but consider them as being members of the joint class of classical radial pulsators. Fig. 5 (sixth row) shows an example of a *Kepler* light curve of a RR Lyrae star along with its amplitude spectrum.

3.7. Solar-like pulsators

Solar-like pulsators (solar class) are intrinsically variable stars showing oscillations driven by turbulent convective motions near their surfaces. Any star with an outer convective zone is expected to show such stochastically excited oscillations. Indeed, following the detection of solar-like oscillations in a number of main-sequence and evolved stars from ground-based data, space-based photometry with the Hubble Space Telescope (HST), WIRE, MOST, SMEI, and in particular CoRoT and *Kepler*, revealed a treasure of pulsational variability in stars with outer convective regions and enabled extraordinary probes of their interiors and improvement of the respective models (see [Hekker & Christensen-Dalsgaard 2017](#) for a review). Stochastically driven solar-like oscillations are well characterized with two global asteroseismic quantities, namely the frequency of maximum power $\nu_{\max}$ and the large frequency separation $\Delta\nu$, which were shown by [Kjeldsen & Bedding \(1995\)](#) to scale with mass, radius, and effective temperature of the star. We do not provide an estimate of the global asteroseismic parameters of solar-like pulsators in our classification scheme, hence no differentiation is made between different evolutionary stages of stars.

Our selection of a sample of solar-like pulsators for the training set is based on the latest release of the APOKASC Catalog ([Pinsonneault et al. 2018](#)). A total of 1800 objects were selected in a random way but making sure each of the targets had a Q9 *Kepler* light curve and oscillations detected with the CAN pipeline ([Kallinger 2019](#)). That being said, our selection of solar-like pulsators for the training set is biased towards red-giant stars with very few main-sequence stars. The majority of those will have a high signal-to-noise ratio detection. An example of the light curve and amplitude spectrum of a solar-like pulsator is shown in Fig. 5 (seventh row).

3.8. Constant stars

Constant stars (constant) are a class of objects that do not show any statistically significant variability on the time scale of 27.4 days. We made a random selection of 1000 objects from the TESS Input Catalog¹² ([Stassun et al. 2019](#)) and simulated their light curves with pure white noise on the 27.4 days *Kepler* time

stamps. The noise level was calculated by adding shot, read, zodiacal and a TESS instrumental baseline noise of 60ppm/ $\sqrt{\text{hour}}$ in quadrature, using the magnitude, effective temperature and galactic coordinates of each object. An example of the light curve and amplitude spectrum is shown in Fig. 5 (last row).

4. METHODS – INDIVIDUAL CLASSIFIERS

We first train four individual classifiers each using different feature sets and learning algorithms. In the next step we then combine these different classifiers using stacked generalization by means of a metaclassifier. The benefit of using this stacked ensemble of classifiers is that we can leverage the individual strengths and weaknesses of each classifier to come to the optimal combination of classifiers and obtain a better predictive performance compared to using just one single classifier.

We constructed the classification framework in a modular way, meaning that the different classifiers can use the same functionality without requiring the use of duplication. We have done this by creating a general `BaseClassifier` class that implements all common functionalities between the different classifiers. The different classifiers then inherit all methods and properties and can define new specific functionalities themselves. This modular set-up makes our framework very flexible and easily allows for additional classifiers to be added later on. As in the other modules of the TASOC pipeline, we make use of Message Passing Interface (MPI) to parallelize our computations. During runtime, all features are also cached in a local SQLite database. In the following subsections we discuss each individual classifier.

4.1. Multiclass *Solar-Like Oscillation Shape Hunter* (*multiSLOSH*)

The multiSLOSH classifier uses image recognition via deep learning to visually determine the presence of the desired signal on a 2D plot of the power density of a star. This is the multiclass generalization of the method described by [Hon et al. \(2018\)](#), where now we classify other types of variability at once instead of only solar-like oscillations. To summarize, a 128×128 binary image of a star's power density spectrum in log-log space is used as input into a 2D deep learning network. The log-log representation of the power density spectrum is used because stars with different types of variability distinctly show different frequency-power profiles in log-log power density spectra. For example, in the case of a solar-like oscillator, one can see the convective granulation background and the Gaussian-like power excess containing the oscillation modes.

¹² <https://tess.mit.edu/science/tess-input-catalogue/>

While the original method has shown to be effective in classifying red giants observed in long-cadence during any amount of time as obtained by TESS (Hon et al. 2018), SLOSH can be very easily generalized towards stars only observed in short-cadence, for example, main sequence, dwarf or subgiant stars. This can be done by modifying the training set that the networks use to learn. To allow for the detection of signals in main sequence or subgiant stars, the plotting range in the 2D image has to be modified. The range in frequency, f_{range} , and power density, P_{range} , for the different evolutionary states are defined by the following:

$$\begin{aligned} f_{\text{range}}(\mu\text{Hz}) &= \begin{cases} [3, 283] & \text{for LC} \\ [40, 4160] & \text{for SC} \end{cases} \\ P_{\text{range}}(\text{ppm}^2 \mu\text{Hz}^{-1}) &= \begin{cases} 3 \times [10^1, 10^7] & \text{for LC} \\ 1 \times [10^{-1}, 10^5] & \text{for SC} \end{cases} \end{aligned} \quad (11)$$

where respectively LC and SC stand for *Kepler* long- and short-cadence data. These ranges are defined in μHz (where one cycle per day (d^{-1}) amounts to $11.57 \mu\text{Hz}$) given that this frequency unit is commonly used in the solar-like community.

The original deep learning implementation from Hon et al. (2018) saved generated plots to image files to be read in later. In this work, we implemented a new method to directly create 128×128 binary array representations of the power density spectra without using a plotting library or input/output to disk. We define 128 even bins in log-log space between the bounds indicated in Eq. 11 that represent image pixels. The default pixel values are one, except for bins that the plotted power spectrum passes through, which take the value of zero. Compared to the original approach, the image arrays that we now generate are computed faster, maintain higher data fidelity, and are better suited for parallel processing.

4.2. *Random Forest General Classification (RFGC)*

The RFGC uses a hybrid self-organising-map (SOM; Kohonen 1990; Brett et al. 2004) and Random Forest (Breiman 2001) classifier, as previously demonstrated on data from the *K2* satellite (Armstrong et al. 2015, 2016). A full methodological description is provided in Armstrong et al. (2016). While the underlying methodology is the same, the features used here have been updated to better account for the new datasets and variability classes considered.

Light curves are initially phase folded, using 64 equal width bins, on the dominant frequency as extracted in

Section 2.1. We also test each light curve using half the dominant frequency, and if the resulting phase-folded light curve shows significantly reduced dispersion, the half-frequency is used. This test ensures the correct value is picked for the orbital frequency of an eclipsing binary, where the presence of primary and secondary eclipses often results in the dominant frequency being double the true binary orbital frequency.

The training set of phase-folded light curves is then used to train a SOM with shape (1,400) using 300 training iterations and a learning rate of 0.1. Training a SOM involves creating a set of template ‘pixels’ which steadily approach similarity to underlying shapes in the input data. In the end the pixels contain representations of various common and uncommon shapes seen in the training set. The index of the closest matching pixel to a test input is then a powerful feature for parameterizing the phase-folded light curve shape.

The actual classification is performed by a Random Forest, implemented through scikit-learn (Pedregosa et al. 2011). The 22 features used are listed in Table 1, including the SOM location described above. We set the parameters of the Random Forest by optimising the out-of-bag score. This led to a Random Forest with 1000 component decision trees, considering a maximum of three features at each node split, with a minimum of two samples required to split an internal node and a maximum tree depth of 15. We use the Gini impurity to measure the quality of a split and in this way select the best splits at the decision tree nodes (Breiman et al. 1984).

4.3. *Supervised random forest variability classifier using high-resolution photometry Attributes in TESS data (SORTING-HAT)*

The SORTING-HAT is a Random Forest classifier with an architecture similar to RFGC. It does not use a SOM, but relies on a set of 13 carefully constructed features in the entropy, Fourier and time domain, as described in Table 1. The use of entropy metrics allows it to differentiate light curves based on their unpredictability and complexity.

The set of hyperparameters is the same as in RFGC, but was independently confirmed by optimising the weighted F_1 score¹³ in an initial version of the classifier, through a general randomized grid search followed by a narrow but complete grid search. This led to a Random Forest with 1000 decision trees, a maximum tree depth of 15, a minimum of two samples required to

¹³ $F_1 = 2 * \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

split an internal node and the usage of the Gini impurity measure.

4.4. *Gradient Boosting General Classification (GBGC)*

Similar to the RFGC discussed in section 4.2, GBGC is a tree-based ensemble method whose trees were constructed with Gradient Boosted Machines (Friedman 2001). In contrast to RFGC, the GBGC is an adaptive method of constructing a model where the classifier aims to correct previous trees in the ensemble by assigning higher weights to the incorrectly predicted samples. The efficiency and generalisation abilities of the GBGC classifier were established using a sample of labelled light curves from the OGLE catalog of variable stars in the LMC (Udalski et al. 2008, 2015) and from the *Kepler/K2* missions by Kgoadi et al. (2019). Eight hyperparameters were adjusted to improve the performance of the classifier. In addition to the number of trees in the ensemble (`n_estimators`) and the optimal depth of the trees (`max_depth`), the fraction of samples in the training set (`subsample`) and features (`colsample_bytree`) during training were also tuned. To ensure convergence was reached in a timely manner, the learning rate of the gradient descent (`learning_rate`) was tuned once the `n_estimators` were determined. Adjustment of the hyperparameters was done to prevent over-fitting and to reduce running time complexities. Optimal hyperparameters were established using a grid search with 10-fold cross validation. This resulted in 500 estimators, a maximum tree depth of 6, a training sample ratio of 0.8, a feature sample ratio of 0.7, and a learning rate of 0.1.

The finalized GBGC classifier was trained on the set of features indicated in Table 1. These were selected using Recursive Feature Elimination with Cross-Validation supplemented by Correlation Based Feature Selection as introduced in Kgoadi et al. (2019). This is a two step feature selection process where recursive feature elimination with cross-validation (Granitto et al. 2006) is applied to select features that best describe light curves and can be mapped to the star classes. To reduce redundancy, the Pearson correlation coefficients were used to remove correlated features from the selected subset through the Correlation Based Feature Selection process (Hall 1999), in which a correlation threshold of 0.65 was applied to remove features. In order to accommodate the class imbalance in our training set, feature selection was done with stratified cross-validation. The GBGC classifier was constructed using XGBoost (Chen & Guestrin 2016) as the base estimator of the model.

5. TESTING AND VALIDATION OF THE INDIVIDUAL CLASSIFIERS

Table 3. Accuracy of each classifier on the training data using 5-fold cross validation and on the holdout set (see Fig. 2 for a graphical explanation of the classifier training and testing procedure). For the Training set (5-fold CV) case we report the mean of the accuracy over each of the five different tested folds. The uncertainty here is equal to standard deviation.

Classifier	Accuracy	
	Training set (5-fold CV)	Holdout set
multiSLOSH	$92.39 \pm 0.89\%$	91.48%
RFGC	$93.41 \pm 0.27\%$	92.56%
SORTING-HAT	$93.79 \pm 0.26\%$	93.70%
GBGC	$93.79 \pm 0.26\%$	91.36%
Meta		94.90%

The individual classifiers are tested and validated in two different ways to ensure that they are not overfitting the training data. For a given training set, we hold out 20% of the data from the start for testing both individual classifiers and the metaclassifier (see Section 6). We partition the remaining 80% of data into five folds or splits of equal size, making sure to include a balanced proportion of all variability classes in each fold. We train on four of these folds and validate on the fifth fold to report one iteration of the performance for all individual classifiers. We repeat this process four more times, but using different folds to train and validate; we are thus cross-validating the individual classifiers over the training set.

Cross-validation is the first approach we use to validate the performance of each classifier. The variance of each classifier over the different folds should be relatively low if they are not overfitting the training set. In Table 3 we report the mean of the accuracy over the five cross-validation folds and report the uncertainty as the standard deviation. The mean scatter of $\sim 0.5\%$ over the cross-validation is due to the small size of some of the classes and the initial training set (0.5% corresponds to ~ 8 stars).

All of the classifiers perform well on the training set, with SORTING-HAT performing best. As we shall see in Section 6, we are not concerned with a single classifier performing better than all the others but more so with the classifiers being uncorrelated with one another. It is important that the individual classifiers have different strengths and each perform best on different parts of the training set if we are to leverage this information in a meta-classification stage.

The second way we validate the individual classifiers is on the 20% initial hold out set that was not used in the previous training and cross-validation step. Whilst

validating the classifiers over cross-validation folds gives a good grasp of how well the classifiers generalize to unseen data, testing the classifiers on a holdout set gives an idea of pure performance and accuracy. We report the accuracy of each classifier in the third column of Table 3.

Overall the holdout set accuracy of the set of classifiers is comparable to their mean accuracy over the cross-validation folds, as for most classifiers the holdout set accuracy lies almost within one standard deviation of the mean cross-validation accuracy. For RFGC alone, we notice that the holdout set accuracy is about 0.6 percent lower than the lower uncertainty bound. In absolute numbers, however, this is still very small and only represents ± 10 out of the 1666 stars in the holdout set. Given that the accuracies on both sets are so similar, we can safely assume that the individual classifiers are fitting the data well and are not overfitting significantly.

We use SHAP (SHapley Additive exPlanations; Lundberg & Lee 2017; Lundberg et al. 2020), a unified approach that connects game theory with local explanations to explain the output of a machine learning model, to compute the feature importance scores. The feature importance plots for SORTING-HAT, RFGC and GBGC are shown in Appendix A including a more detailed explanation. For RFGC we find that the zero-crossing parameter and point-to-point differences are the most important features, while for SORTING-HAT it is clear that the multiscale entropy (MSE) together with the first fundamental frequency and skewness are the most important attributes in the classification process. Lastly, for GBGC the variability index is by far the most important. We do not plot the feature importance scores for multiSLOSH given that it is a neural network classifier that does not rely on a set of predefined features, but rather learns a set of weights that define the importance of each region in the power density spectrum image.

6. THE METAClassIFIER

Each of the individual classifiers described in Section 4 predicts the class probability scores for each light curve. We combine the predictions from this ensemble of classifiers using stacked generalization (Wolpert 1992), in which we turn to a metaclassifier that takes the probabilities outputted by the individual classifiers as its features to produce overall class probabilities for each light curve. This metaclassifier accounts for the relative strengths of the individual strong classifiers in the ensemble (see Schapire 1990 for a description of strong versus weak).

6.1. Training the Metaclassifier

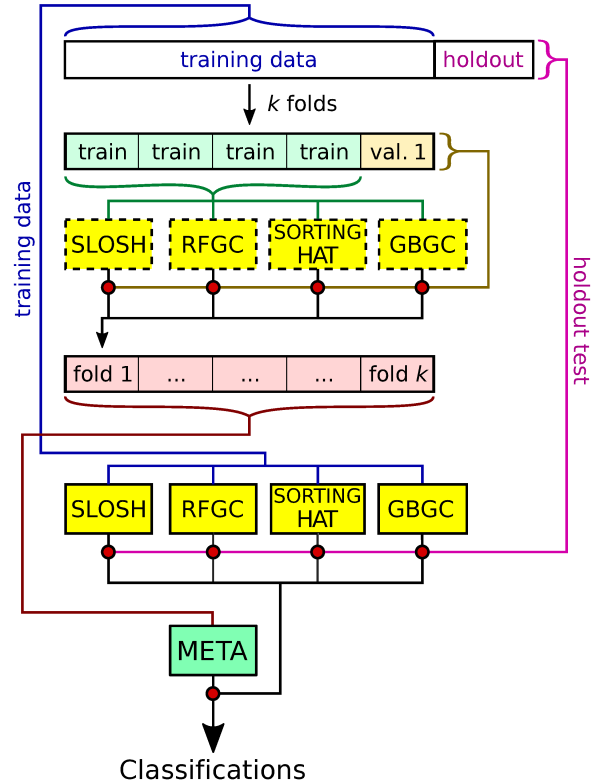


Figure 6. Graphical representation of the classifier training and testing procedure. 80% of the data set is split into k stratified folds for cross-validation, where $k = 5$. Class probabilities for data in each fold are predicted by the supervised individual classifiers trained on the other $k - 1$ folds. The training class probabilities from each individual classifier are used to train the metaclassifier. The individual classifiers used to characterize the unseen data are trained on all of the training data. The success of the overall classification is tested by classifying the holdout data with the individual classifiers, and then by using their predictions as input into the metaclassifier.

The stacked nature of our overall classification scheme could lead to overfitting and poor generalization to the unseen TESS data if the classifier is not trained carefully. Our approach to training the metaclassifier and individual classifiers together closely follows Algorithm 19.7 in Aggarwal (2014). We represent our application of this training algorithm graphically in Figure 6.¹⁴

As explained in Section 5, for a given training set, we take 20% of the data from the start as a holdout set to test the trained ensemble of individual classifiers. We split the remaining data set into k folds and produce class probabilities for each by cross validation. We predict the class probabilities for each fold using the classi-

¹⁴ Inspired by the illustration at http://rasbt.github.io/mlxtend/user_guide/classifier/StackingCVClassifier/

True Class	constant							100
	aperiodic	1				1	98	
	contactEB/spots	<1			<1	5	92	<1
	gDor/spB				5	90	5	
	dSct/bCep			<1	97	2	<1	
	RRLyr/Ceph			92				8
	transit/eclipse		96				4	
	solar	94				<1	<1	5
		solar	transit/eclipse	RRLyr/Ceph	dSct/bCep	gDor/spB	contactEB/spots	aperiodic
Predicted Class								

Figure 7. Normalized confusion matrix of the metaclassifier for the holdout set in percentages. Each element shows the fraction of stars that were predicted as positive for a particular class (column) over the total number of stars that truly belong to that class (row). The diagonal shows the fraction of stars that the classifier correctly predicted as positive for that class (i.e. the recall rate = $\frac{TP}{TP+FN}$, where TP is the number of True Positives and FN the number of False Negatives).

fiers that are trained on the other $k-1$ folds. We assume that the performance of the classifiers trained across k folds approximates the performance of the models trained on all of the training data. The cross-validated class probabilities from each of the individual classifiers on the training data are the inputs used to train the metaclassifier. The performance of the metaclassifier is finally tested on the holdout data by using the holdout set class probabilities predicted by the individual classifiers trained on the training data (indicated in blue on Fig. 6) as input.

The algorithm we use for the metaclassifier is a Random Forest with a similar architecture to RFGC (see Section 4.2), but with the number of estimators and maximum tree depth constrained to respectively 100 and 7, to avoid overfitting. This is chosen over a simpler scheme such as majority/soft voting because we want to leverage the potential correlations between classes. The meta classifier, like the individual classifiers, predicts the class probabilities per star. We note that the class

probabilities predicted by the metaclassifier are well calibrated, but not perfect. It might thus be better to interpret them as a ranked confidence score rather than in a purely probabilistic fashion. If the metaclassifier assigns a confidence of 0.8 to 100 predictions, we should not expect that exactly 80 of those are correct. However, if we have a star with a confidence of 0.3 and a star with a confidence of 0.7, we can safely assume that the second one has a much higher probability of belonging to the class than the first one.

6.2. Metaclassifier Testing and Validation

The metaclassifier obtains an accuracy of 94.90%, a substantial improvement over any single individual classifier. However, given that there are class imbalances in our training set the overall accuracy only provides a limited amount of information. We therefore also look at the confusion matrix as this gives a more detailed view on the metaclassifier’s performance per class. The confusion matrix for the metaclassifier is shown in Fig. 7. The classification rates (or recall scores) per class range

from 90% for the γ Dor/SPB class to a near perfect score for the constant class. A detailed look reveals that the lower score for γ Dor/SPB is mostly caused by confusion with the δ Sct/ β Cep and contactEB/spots class. Our visual analysis shows that the former can be explained by the presence of hybrid pulsators in the training sample, while the latter is caused by γ Dor/SPBs containing either some rotational signal or low frequencies that resemble those of contactEB/spots. We also notice some confusion between the aperiodic and contactEB/spots classes. This is mostly caused by the fact that both classes can mimic each other on the short time scale of 27.4 days. Lastly, there is a fraction of solar-like oscillators being predicted as aperiodic variables, where we find that they all have low $\nu_{\max}$ values, hence their light curve and power spectrum properties are similar to those of aperiodic stars. The high percentage for RRLyr/Ceph misclassifications is due to the small class size and in absolute numbers only concerns one star.

In Fig 8 we show the feature importance plot for the metaclassifier, which allows us to analyze the contribution of each individual classifier towards the final prediction. The hatched regions indicate the most important feature for a specific class here. This reveals that the multiSLOSH_solar probability is the most important feature in the classification of solar-like oscillators, followed by SORTING-HAT. This could be expected given that SLOSH was initially designed to classify this type of star and in the case of SORTING-HAT the entropy features allow it to capture the stochastic nature of the signal. The same order holds for the gDor/SPB class. GBGC is most the important classifier in classifying transit/eclipse signals while it is RFGC for the aperiodic, RRLyr/Ceph, constant and dSct/bCep classes. SORTING-HAT is the primary classifier for contactEB/spots stars. It is interesting, however, that in the contactEB/spots class, SORTING-HAT is followed by the multiSLOSH probability of being a solar-like star. By plotting the SHAP values of every feature for every star, specifically for the solar class, we analyze the impact of each feature on the model output (i.e. the probability of being classified as solar). This reveals that a high multiSLOSH_solar probability lowers the predicted probability of being a contactEB/spots star and vice-versa. The feature importance plot in Fig. 8 clearly shows that the metaclassifier’s strength is in combining the different individual classifier results.

We also assess performance by looking at the receiver-operator characteristic (ROC) curves. The ROC curves illustrate the diagnostic ability of a classifier by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) for different classification thresholds. This

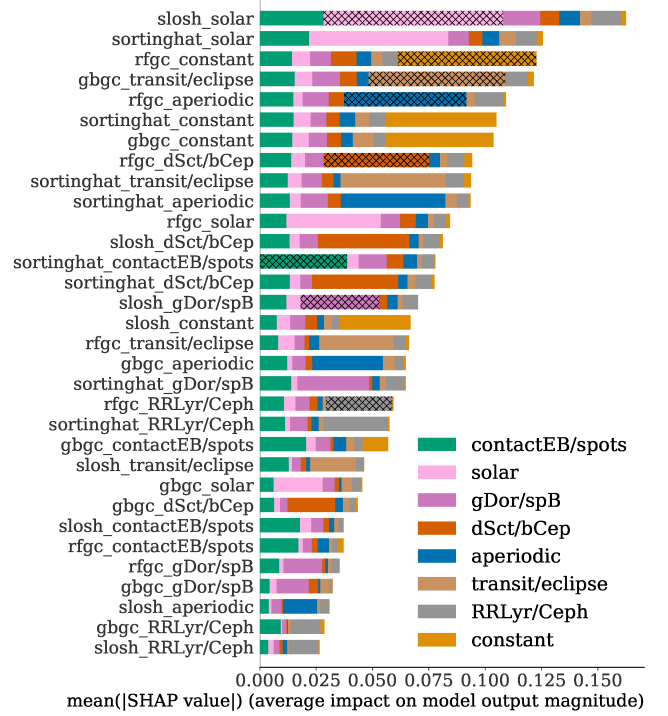


Figure 8. Metaclassifier feature importance from SHAP. The hatched regions indicate the most important feature per class.

allows us to assess the performance of the classifier at each threshold. We calculate the ROC curves for each class using a one-vs-rest methodology (Fawcett 2006). The ROC curves per class are shown in Fig. 9. Ideally, the ROC curve should be as close to the top left corner (0,1) as possible, because for this threshold on the curve the classifier is making a high number of correct classifications with a small amount of false positives.

Final class labels are commonly assigned to the class with the highest probability, which is equivalent to using a probability threshold of $1/C$, where C is the number of classes. In case more than one of the predicted class probabilities of a star exceeds its respective threshold, the star is assigned to the class with the highest probability. When dealing with class imbalance, however, this $1/C$ approach often does not lead to the optimal results (Provost 2000). We therefore opt to fine-tune the classification threshold by choosing for each class the threshold that maximizes the TPR and minimizes the FPR, which is the point on the ROC curve that is closest to the top left corner. This point can be determined by finding the threshold that maximizes Youden’s J statistic (Youden 1950), which is the difference between the TPR and FPR. Given that we have one ROC curve per class, this implies that we also have a different threshold for each, reflecting the classifier’s differing ability in

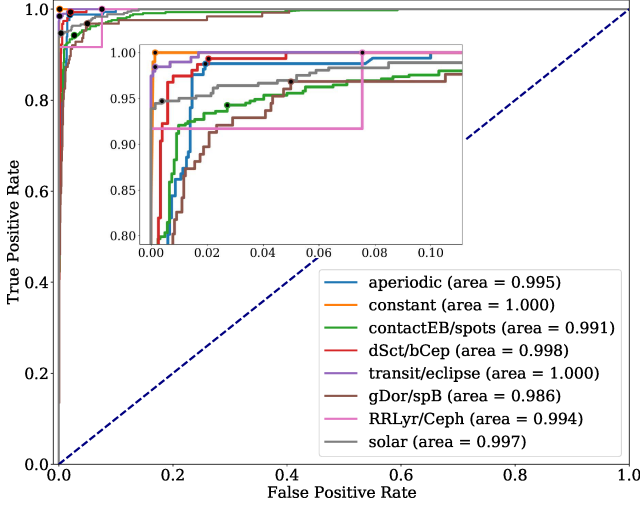


Figure 9. Receiver-Operator Characteristic (ROC) curves of the metaclassifier for each class. The Area Under the ROC (AUROC) curve is indicated next to the classes in the legend and the dots represent the TPR and FPR for the chosen probability thresholds. The dashed line indicates the ‘random chance’ curve.

Table 4. Classification thresholds per class.

Class label	Probability threshold
aperiodic	0.211
constant	0.593
contactEB/spots	0.262
dSct/bCep	0.049
transit/eclipse	0.157
gDor/SPB	0.135
RR Lyr/Ceph	0.101
solar	0.296

identifying the class members of each variability class. The obtained thresholds per class are given in Table 4.

As an aggregate performance measure across all probability thresholds used in the ROC curve, we can measure the Area Under the ROC curve (AUROC). The AUROC represents the probability that the classifier assigns a higher probability to a random positive example than to a random negative example. It is thus a measure of how well the classifier predicts the correct class. Given that we are working with a one-vs-rest methodology, it means that the respective ROC class is the positive class and that the other classes belong to the negative class. The confusion between the gDor/SPB and contactEB/spots class causes their AUROC values to be slightly lower compared to the other classes.

7. VALIDATION ON FULL *KEPLER* Q9 DATA SET

Table 5. *Kepler* Q9 classification summary: number of stars per class for each thresholding method.

Class label	# stars per threshold type	
	1/C	Youden’s J
aperiodic	3 711	3 711
constant	5 061	0
contactEB/spots	140 566	139 059
dSct/bCep	1 758	1 758
transit/eclipse	1 563	1 563
gDor/SPB	2 263	2 263
RR Lyr/Ceph	96	96
solar	12 225	12 185
<i>unknown</i>		6 608

We validate our classification scheme by applying it to all 167 243 stars observed in *Kepler* Q9, but with their light curves cut to the first 27.4 days. We start with the default methodology as described in the previous sections, then test the effect of linear detrending, and ultimately assess the advantage of introducing an instrumental class. For each of those additional scenarios, we assess both the results on the holdout set and on the *Kepler* Q9 data set. The assessment is achieved by the summary statistics of both data sets and by visually inspecting random sub-samples of 1000 light curves in each class for the *Kepler* Q9 classification results.¹⁵

7.1. The default scenario

The results obtained by applying our framework to *Kepler* Q9 data are summarized in Table 5. The left column lists the numbers for the label assignments being made according to the highest probability, while the right column gives those according to the optimized probability thresholds. Overall, we see that all predicted classes, apart from contactEB/spots, have classification rates similar to those in the confusion matrix in Fig. 7. The high number of stars in the contactEB/spots class can be explained by the fact that light curves that are not assigned to any of the other classes, for example with a dominant instrumental signal in the low-frequency domain, end up in this bin. A careful visual inspection of the light curves and amplitude spectra of 1000 random sub-samples in each class strengthens the above conclusion; below we present a concise summary of our visual analysis.

The aperiodic variables are identified robustly by our methodology with an overall low number of misclassifications. After inspecting 1000 light curves randomly

¹⁵ Before taking the random sample we first removed the stars that were included in the training set.

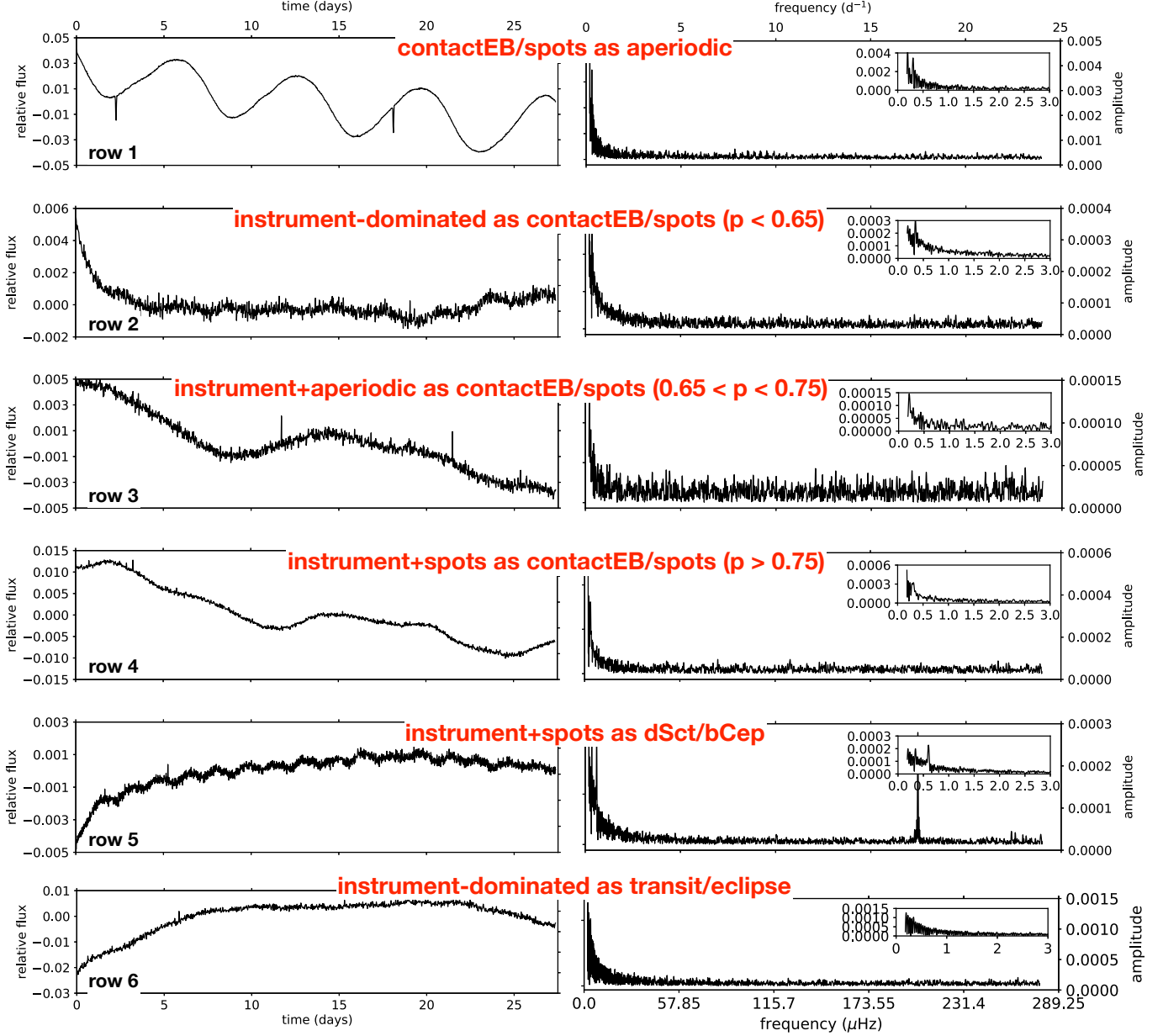


Figure 10. Examples of the misclassified *Kepler* Q9 light curves. Left and right columns show the light curves and the amplitude spectra, respectively. Note the different scale on the Y-axis of the plots.

selected from the respective class, we confirm that some 97% of the light curves indeed exhibit aperiodic type variability as demonstrated in the first row in Fig. 5. The most common misclassifications (about 3% in total) belong to the contactEB/spots class and are the light curves resembling rotational modulation and/or binary ellipsoidal signals, in many cases with the coverage of a single rotation/orbital period. The median probability value for misclassifications is found to be $p(x) \approx 0.50$. The worst-case scenario misclassification light curve is shown in Fig. 10 (first row) where likely a close eclipsing binary got (mis)classified as an aperiodic variable.

The contactEB/spots class suffers the most from misclassifications and partially resembles properties of miscellaneous classes often employed by other light curve classification methods (e.g., [Debusscher et al. 2011](#)). Fig. 11 (orange line) shows the probability density function for the contactEB/spots class. We immediately notice an excess of objects in the low probability regime ($p(x) \lesssim 0.55$) as well as a double-peak feature at high probabilities ($p(x) \gtrsim 0.65$). Owing to this distribution, we divide the contactEB/spots class into three probability bins and visually inspect 500 randomly selected light curves in each of them: *i*) $p(x) < 0.65$, *ii*)

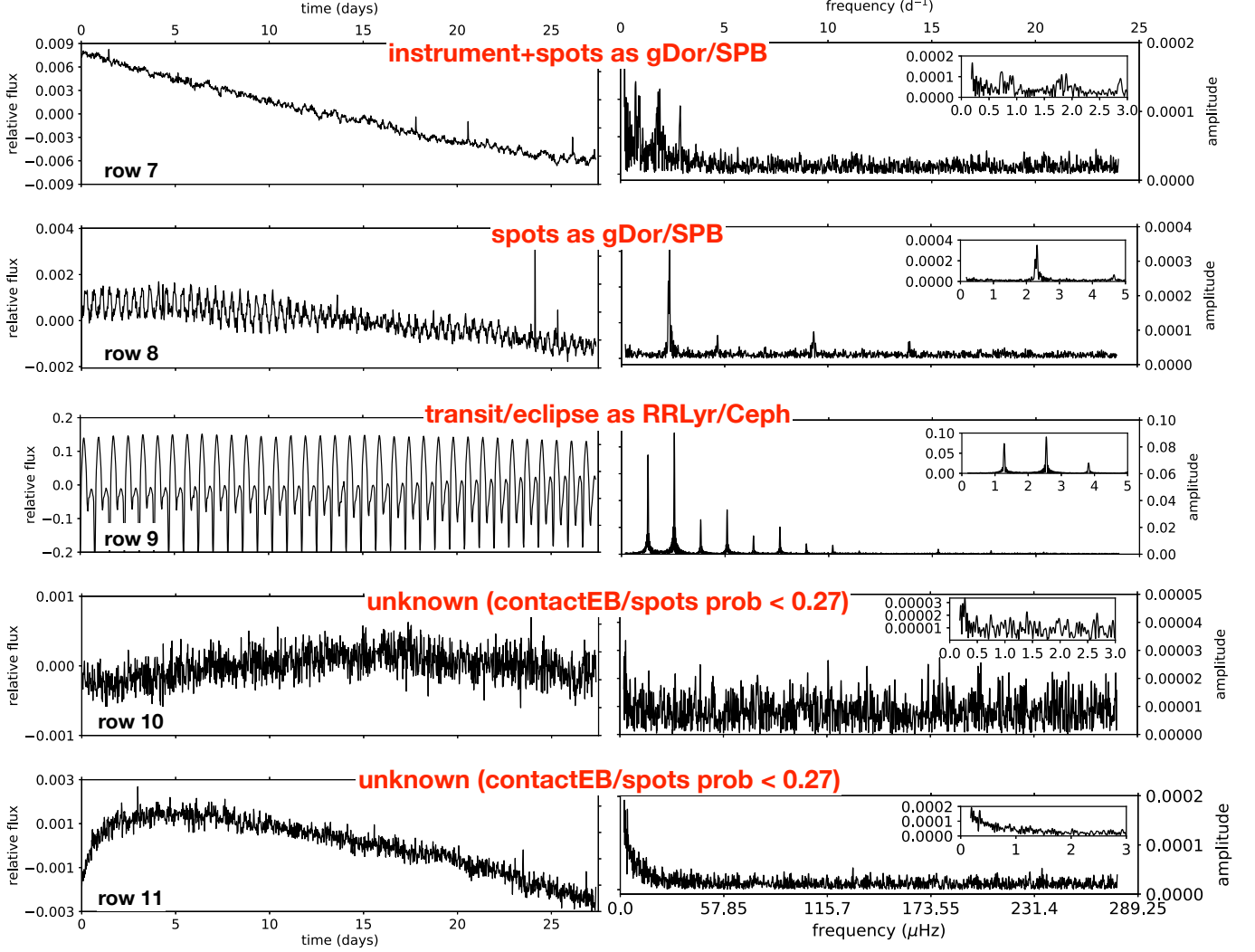


Figure 10. Continued

$0.65 < p(x) < 0.75$, and *iii* $p(x) > 0.75$. We find that the lowest probability bin ($p(x) < 0.65$) contains some 97% misclassifications. Among those the dominant fraction (about 90%) are light curves that exhibit some sort of instrumental signal (see the second row in Fig. 10). This can be either truly instrumental in origin or due to inferior data processing. The intermediate-probability bin that is associated with the first peak in the kernel-density plot ($0.65 < p(x) < 0.75$, Fig. 11) is also found to be rich in misclassifications (overall about 92%). Yet, the major difference with the low-probability regime is that the fraction of light curves that exhibit pure instrumental signal is significantly lower, at around 55%. In the rest of the light curves, the instrumental and the true astrophysical signals are found to co-exist, as illustrated in the third row in Fig. 10. In this particular example, a weak astrophysical aperiodic signal (on the time scale of 27.4 days) co-exists with low-frequency signal due to inferior data processing. Lastly, the highest probability

bin associated with the tallest peak in the probability density function ($p(x) > 0.75$, Fig. 11) contains 16% misclassifications that are pure instrumental in origin. About 34% show both instrumental and astrophysical signals. We note that the latter are not necessarily misclassifications, it is just that we visually identify the instrumental signal as being the dominant one in the respective light curves (the fourth row in Fig. 10). Finally, we note that pure astrophysical misclassifications are dominated by aperiodic variables and is at the level of some 17%. Many of those seemingly aperiodic signals might in fact be rotational variables with periods longer than 13.7 days and therefore spanning less than half of the rotation cycle. Hence they are visually classified as aperiodic stars. That said, we recommend a probability threshold of $p(x) \gtrsim 0.75$ for the high-confidence selection of contactEB/spots variables for this class (an example is shown in the second row in Fig. 5). This threshold deviates from the one calculated using Youden's J statistic

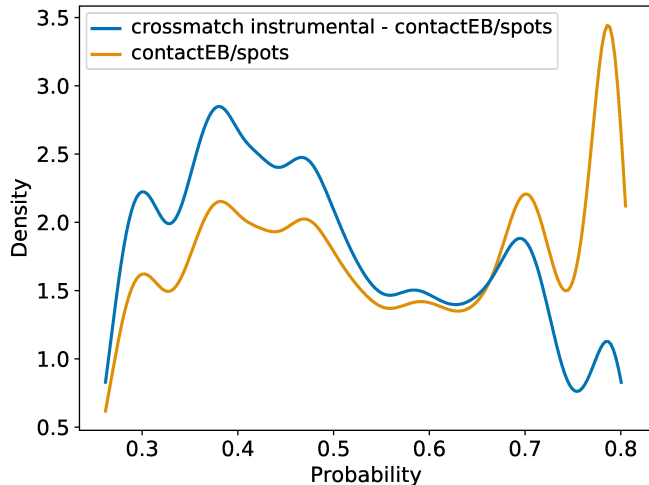


Figure 11. Kernel density estimate (KDE) plot comparing $p_{\text{contactEB/spots}}^{\text{default scenario}}(x)$ of the instrumental class cross-matched to the original (i.e. from the default set-up) contactEB/spots class, against the complete original contactEB/spots class, with class assignments based on Youden’s J statistic (see Sect. 6.2 for a description of Youden’s J). Stars from the training set have been subtracted from both sets.

because our training set is largely free from any instrumental signal, while this type of signal is present in the full *Kepler* Q9 data (we will elaborate on this point in sect. 7.2), resulting in a suboptimal contactEB/spots threshold. The results with probabilities $p(x) \lesssim 0.75$ should be taken with caution and one should keep in mind that the number of genuine astrophysical signals in this particular class drops substantially towards low probability values.

The dSct/bCep variables are identified with high confidence by our methodology, where the overall fraction of misclassifications amounts to some 3.5%. The vast majority of misclassifications are due to spurious frequencies found in the high-frequency domain (the fifth row in Fig. 10). We consider those frequencies as spurious because exactly the same frequencies are found in multiple objects indicating their non-astrophysical origin. At the same time, we did not find any indication of these particular frequencies being instrumental in nature as those are not listed as such the latest *Kepler* Data Release Notes. The median probability value for the misclassified light curves is $p(x) \approx 0.45$. A considerable fraction of the identified δ Sct stars also exhibit low-frequency variability (either due to g-mode oscillations or rotational modulation), yet the high-frequency component is significant in all the detections and is the dominant one in the majority of them.

The transit/eclipse class is among the cleanest identified with our method, containing about 10% misclassifications overall. All misclassifications look alike and are

due to imperfections in the data processing mimicking a flux drop in the light curve, most often at the beginning/end of the dataset. A typical example of the transit/eclipse misclassification is shown in the sixth row in Fig. 10. We also note that the type of light curve shown in the fifth row in Fig. 10 has high chances of being misclassified as a transit/eclipse variable, in absence of the high-frequency peak. We find the median probability value for the misclassifications to be $p(x) \approx 0.45$.

The gDor/SPB class suffers from about 30% misclassifications, either from astrophysical signal of different origin or from low-frequency signal due to imperfections of data processing. The median probability value for misclassifications appears to be $p(x) \approx 0.52$ and the most common astrophysical misclassifications are stars that belong to the contactEB/spots class. We show a typical example of a misclassified light curve in the seventh and eighth rows in Fig. 10. The Fourier transform of the light curve reveals a rich variability spectrum at low frequencies, possibly with a harmonic structure. Owing to the frequency range of gravity-mode oscillations observed in γ Dor/SPB stars and to the short time span of light curves that are being classified (see the fifth row in Fig. 5), contactEB/spots class members are indeed the primary candidates for an astrophysical misclassification in the gDor/SPB class. We also note that the particular example shown in the seventh row in Fig. 10 might not be the actual misclassification but is identified by us visually as such because of a short duration of the respective light curve.

The RRlyr/Ceph class of classical pulsators is found to be small (see Table 5) which is expected given the location of the *Kepler* field. Misclassifications amount to about 60%, are mostly astrophysical in origin, and are dominated by contactEB/spots or transit/eclipse class members. All three types of objects (including RRlyr/Ceph) have their dominant signals in the low-frequency domain and will often show a harmonic structure. However, owing to the characteristic shapes of their light curves (as shown in the sixth row of Fig. 5), RRlyr/Ceph stars are usually readily distinguished from other classes of variable stars. Indeed, only a small fraction of binaries and/or rotational variables with light curves that closely resemble those of the RRlyr/Ceph class are expected to exist and hence get misclassified to the class of classical pulsators. The high relative number of binaries and rotational variables that we find in this class is very likely the result of the contactEB/spots and transit/eclipse classes being at least two orders of magnitude larger than the RRlyr/Ceph class itself, hence increasing the chance of misclassification. An example of the light curve and amplitude spec-

trum of an eclipsing binary misclassified as RRLyr/Ceph is shown in the ninth row in Fig. 10.

The solar class is the second largest. The number of misclassifications is found to be small in this class (about 4%) and is mostly instrumental in origin. The median probability value for misclassifications is found to be $p(x) \approx 0.55$.

The unknown class contains objects that do not satisfy the Youden's J statistics-based thresholds per class as listed in Table 4 (see Sect. 6.2 for a description). By comparing the class sizes before and after applying the thresholds (see Table 5) we notice that the unknown class comprises the entire class of constant stars as well as a small fraction of the lowest probability objects from the contactEB/spots class. The constant class gets marked as unknown due to the fact that the calculated probability threshold of 0.593 is very high. This happens because the classifier achieves a near perfect classification rate for the constant class (see Fig. 7) and is thus very confident in classifying stars as such. Therefore, during testing, stars are either predicted not belong to the constant class at all (i.e. $p(x) \approx 0$) or they are confidently classified as constant. In the latter case, the lowest probability (which is still high in absolute terms) of a star that is being classified as constant is set as the threshold (due to the mathematical calculation). In reality, however, it appears that there are no stars as distinctly constant as in our training set. This makes sense given that this class is simulated in the training set while existence of constant stars is not assured. 99% of objects found in the unknown class do not show any significant astrophysical signal, with two typical examples being shown in the two bottom rows (10 and 11) in Fig. 10.

7.2. Additional classification set-ups

We also test the effect of automatically removing a linear trend from all *Kepler* 27.4 days light curves prior to computing the Fourier- and time-domain features. The results on the holdout set and *Kepler* Q9 show no performance gain over the default set-up, hence we stick to using the original light curves. We do not test detrending with higher degree polynomials because this can have undesirable effects on the classification as *i*) the original light curves may be significantly distorted, and *ii*) the signal of long-period variables may be largely filtered out during the process.

One of the key findings from running our default set-up is that the contactEB/spots class is largely overpopulated, with a clear tendency for a large number of misclassifications towards the low probability values by light curves containing some sort of an instrumental signal.

We note that we use the term “instrumental signal” to mark a signal that is either truly instrumental in origin or is the result of sub-optimal detrending/correction of the data. To overcome the above-mentioned drawback, we opt to introduce an instrumental class with properties resembling those of light curves affected by the instrumental trends and/or sub-optimal data processing. We use a sub-sample of the contactEB/spots class light curves whose probability values were found to be of $p(x) \lesssim 0.65$ to manually select a training set for the instrumental class based on the visual inspection of the light curves. To preserve the balance with other variability classes in the training set a total of about 1100 light curves are selected.

The most notable differences after introducing the instrumental class are *i*) a considerable reduction of the size of the contactEB/spots variability class by about a factor 3.5, and *ii*) a much smaller size of the unknown class. This happens because the originally low-probability (lower than the respective thresholds reported in Table 4) objects in the various classes are now classified with high confidence as members of the newly introduced instrumental class. Hence there are considerably less candidates to feed the unknown class in the latter.

Furthermore, we cross-match the newly obtained instrumental class with the original contactEB/spots class and find about 70% of overlap. The probability density plot for the cross-matched sample is shown in Fig. 11 (blue line) where the distribution is evidently skewed towards low probabilities. Therefore, we conclude that introducing an instrumental class does not necessarily improve the overall performance of the method, instead a considerable fraction of light curves that receive low confidence values in their respective classes are moved to the class of instrumental variables.

While not having clear advantages, the disadvantages of introducing an instrumental class are that it is not only instrument-dependent but that it is also extremely sensitive to the way data from a given instrument are being processed. Therefore, an instrumental class proves impractical as it has to be re-designed each time the data from a given instrument are being reprocessed and/or the methodology is being applied to data from a different instrument. A much more practical solution is the one outlined and employed in Sect. 7.1, i.e. a recommended probability-based threshold to separate high-confidence detections of genuine contactEB/spots variables from their low-confidence counterparts that are most likely not astrophysical in origin.

7.3. Effects of photon noise

The level of recognizable astrophysical signal in a light curve is tightly related to the amount of photon noise present in it, which depends on the stellar magnitude. Given that we did not exert any particular influence on the distribution of the magnitudes in the training set, we test the sensitivity of the metaclassifier with respect to increasing photon noise. As for the relation between magnitude and noise level, we base ourselves on the 10th percentile RMS CDPP (Combined Differential Photometric Precision) measurements presented in the TESS Data Release Notes: Sector 5, DR7. We multiply the values by $\sqrt{2}$ to account for the 30-min sampling.

For each class in the holdout set, we select the 20 stars with the highest probability and remove all stars with *Kepler* magnitude < 15 . We leave out the constant stars given that they are simulated white noise already. We then add noise to the sampled light curves in steps of 0.5 magnitude, with the maximum number of steps restricted to 8, which is equivalent to a magnitude 4 increase. The added noise is Gaussian with mean zero and standard deviation equal to 30-min CDPP value for the desired magnitude. The new equivalent magnitude is constrained to be brighter than 15.5. If this is reached before the maximum number of allowed steps, no further noise additions are done for the star. We choose this constraint because the T'DA Photometry pipeline (Handberg et al. 2021) will process TESS stars down to magnitude 15.

Once we have calculated the new noisier light curves, we classify those in each step with the metaclassifier and analyze how the overall predictions change with added photon noise. Fig. 15 shows how stars move between the different classes when more noise is added to their light curves. The relatively brightest stars are shown on the left and relatively faintest on the right. The colors of the streams indicate the true variability class, and the bars indicate the predicted class. The height of the bars corresponds to the number of stars in that bin. The number of stars decreases from left to right because stars are eliminated once their new equivalent magnitude exceeds the 15.5 threshold. In Fig. 16 we show how the magnitudes evolve over the different steps. We start with 115 stars on the left and end up with 36 in the rightmost bin.

We can see from Fig. 15 that when the noise level increases (i) the majority of solar-like stars get classified as constant (ii) a significant fraction of contactEB/spots star get classified as constant and (iii) most aperiodic stars end up being classified as contactEB/spots. The reason for (i) is physical in origin and results from the fact that the added noise is much larger than the oscillations in the original light curve, causing the new

light curves to be dominated by white noise and get classified as such. The solar class is also the most varied one in terms of time scales and so the location of the oscillations dictates to which bin a star moves into when adding noise, causing some of them to be classified as other variability types as well. In a similar manner, we can see that the contactEB/spots stars that get classified as constant (ii) are actually cool and spotted stars in which the noise becomes larger than their oscillations. Only the hot and chemically peculiar stars with high amplitude variability that is stable on longer time scales survive. The reason for (iii) can be attributed to a training set bias and occurs because aperiodic and contactEB/spots stars can mimic each other on 27.4 day time scales, and because in our training set the contactEB/spots class tends to be more noisy than its aperiodic counterpart, causing these stars to be classified as such.

When we connect these findings to the magnitude distribution of our full training set (Fig. 17), we conclude that one should be careful when interpreting the predicted probabilities of stars that do not lie within the magnitude range of the training set. More specifically, a decreasing magnitude for stars that are part of the solar, contactEB/spots or aperiodic class corresponds to an increasing uncertainty over their probabilities. Hence, when interpreting the results, it is important to, in addition to the assigned probabilities, also look at the magnitude of the target. If the magnitude is much fainter than those of the training samples and it belongs to one of these three classes, caution should be paid when interpreting the results. For other classes, such as transit/eclipse, this effect is not present because the amplitude of the signal is often much larger than the added noise. It is thus important to note that a bright star, even in the case of (i), (ii) and (iii), does not always mean that there is a very clear signal while a faint stars does not necessarily mean we have an indistinguishable signal. It is the amplitude of the signal relative to the noise that matters.

8. DISCUSSION AND CONCLUSIONS

8.1. Summary and Discussion

The TESS Data for Asteroseismology pipeline is designed for a largely automated processing and high-level interpretation of TESS space-based photometric data. As depicted in Fig. 1, the first two modules of the pipeline are designed for the extraction of light curves from the TESS Full Frame Images (Handberg et al. 2021) and for their subsequent optimal correction for systematic effects (Lund et al. in prep.). In this work, we have designed a third module of the T'DA pipeline

that performs an automated classification of the corrected light curves according to their type of variability.

We combine four individual classifiers into a meta-classifier using stacked generalization. Out of the four individual classifiers, RFGC (Armstrong et al. 2016, Sect. 4.2) and SLOSH (Hon et al. 2018, Sect. 4.1) have been previously published, while SORTING-HAT (Sect. 4.3) and GBGC (Sect. 4.4) were additionally developed to enhance the T'DA pipeline classification module appreciably. We show that by stacking the predictions of this set of different individual classifiers we obtain a substantial improvement over any single one, because the metaclassifier is able to learn their relative strengths. We are able accurately classify light curves according to their general variability type, without relying on any extra information other than the light curves.

Although inspired by the amount of TESS data currently collected, our ultimate goal is to design an automated pipeline for the end-to-end processing of high-cadence and duty-cycle space-based photometric data, irrespective of whether these come from the retired CoRoT and *Kepler*/K2 missions, currently operational TESS mission, or future space-missions such as PLATO (Rauer et al. 2014). Hence, in this work, we make use of the *Kepler* mission legacy, both in terms of the available high precision, cadence, and duty-cycle data and the published catalogs of variable stars, to train, validate and test our classifiers. The training set is carefully built from the existing catalogs with a subsequent vetting of light curves in all eight variability classes used in our classification scheme. All individual classifiers as well as the metaclassifier are trained on 80% of the compiled training set, while the remaining 20% are kept as a hold-out set to test and validate the method. We obtain an overall accuracy of 94.9% on the holdout set with some small differences between the different classes.

We further apply the designed classification scheme to the *Kepler* Q9 data set that has been truncated into 27.4 days light curves. In addition to testing our default classification set-up, we also test the effect of linear detrending and the introduction of an extra instrumental class to isolate light curves dominated by the instrumental signal. We show that although the latter allows for a significantly lower number of misclassifications of the sub-optimally processed light curves in some of the classes, it has the disadvantage that the instrumental class has to be re-designed each time the method is applied to the re-processed data from the same space-mission and/or data obtained by another mission.

Given that we currently use 27.4 days light curves, one of the expected and detected (astro)physical limitations of our method are apparent misclassifications of objects

whose variability on a 27.4 days time scale does not necessarily resemble their true origin. A common example is non-resolved rotational variability in cool stars that gives rise to an overdensity of low frequencies in the Fourier transform of the light curve causing a confusion with the class of g-mode pulsators and/or aperiodic variables. Another example is the flux drop in a light curve due to sub-optimal data processing which mimics a single transit/eclipse event in the time-domain and gives rise to a misclassification as a transiting/eclipsing object. Other than that, we find that the predicted classes have classification scores similar to those in the confusion matrix based on the hold-out validation set (see Fig. 7).

Generalizing our framework to TESS will still require adjustments because we are currently training our classifiers on *Kepler* data. Not only does *Kepler* have a different underlying distribution compared to TESS, possibly requiring domain adaptation techniques, TESS also has a worse photometric precision (and hence more noise), more blending, and more systematics that we cannot characterize very well yet. That said, there is no one-to-one correlation between the results obtained based on *Kepler* data in this work and the expectations for TESS data. In order words, we cannot simply extrapolate the results of the performance of our classifiers to TESS data prior to exploring domain adaptation, performing initial classification of TESS dataset, and ultimately (re)training based on the actual TESS data. We note, however, that the performance will not necessarily drop when transitioning to TESS data, it can also be as high or higher than in this work.

We make both the methodology and the results of its application to the *Kepler* Q9 27.4 days data using the default set-up publicly available to the community. Our training set, individual classifiers, and the metaclassifier can be accessed through the dedicated GitHub repository¹⁶ as well as through the TESS Asteroseismic Consortium (TASOC) Wiki pages¹⁷. The predicted class probabilities and class labels for the *Kepler* Q9 27.4 days are released in electronic format; a snippet of the class probabilities table is shown in the Appendix (Table. 6).

8.2. Future prospects

With the machinery built, our immediate future prospects include:

- Classification of all *Kepler* stars based on *i*) 1-year data to mimic TESS Continuous Viewing Zone

¹⁶ <https://github.com/tasoc/starclass>

¹⁷ <https://tasoc.dk/tda/>

(CVZ) operations and enable direct comparison with the results presented in this work; *ii*) 2-year data to mimic PLATO Long Pointing Field (LPF) operations enabling an important set of tests for the PLATO Consortium; and *iii*) 4-year data to provide a full *Kepler* classification catalogue and quantitatively assess performance of our method on ultra-high precision data by cross-matching with the existing *Kepler* catalogues. At this step, we will consider using extra information, such as photometric colours, Gaia parallaxes, etc., in order to break the existing degeneracies within and between the individual variability classes. This particular step covers our intended “second-level classification” (as depicted in Fig. 2) where we aim to distinguish between different evolutionary states of solar-like pulsators (dwarfs vs. sub-giants, RGB stars vs. red-clump stars), between sub-groups of g - (γ Dor vs. SPB variables) and p-mode (δ Sct vs. β Cep stars) pulsators, etc.

- Inclusion of a learning algorithm capable of identifying transient phenomena, such as stellar flares, Be star outbursts, etc. For this, we will consider existing algorithms such as STELLA¹⁸ (Feinstein et al. 2020) which will be adapted to the needs of our metaclassifier similarly to the RFGC and multiSLOSH methods.
- Inclusion of an unsupervised learning algorithm to help identify misclassifications and search for over-densities in the feature space within the identified supervised classification module variability classes. This particular step is depicted in Fig. 2 as the “unsupervised methods” box and will strengthen our classification scheme by allowing for the detection of additional variability (sub)classes.
- Inclusion of statistical features for an improved classification of aperiodic autocorrelated signals. For this, we will consider tests such as the Durbin-Watson statistic for serial autocorrelation and the Kullback-Leibler divergence to measure the disparity against white noise.
- Transition to TESS data that are processed with the corresponding T'DA pipeline light curve extraction and systematics correction modules. At this step, we also envision an iteration between all three modules of the T'DA pipeline, in particular to inform the light curve correction algorithms

on the variability time-scales that should be preserved rather than removed for specific classes of objects. In terms of the corresponding data releases, we plan them jointly with the light curves themselves on a per sector basis and will make our results publicly available through the MAST and TASOC databases. The accompanying TESS classification papers for the nominal missions are also foreseen and will be based on the full year of TESS data, i.e., per TESS observational hemisphere.

- Integration of the variability catalog into the TASOC database search-interface¹⁹. This interface will allow for a quick and convenient search of variable stars according to user-defined selection criteria. As concrete examples, one will be able to opt for an all-sky search of δ Sct variables that have been identified with a user-defined confidence with our classifiers, or stars classified with probability in multiple classes (e.g. both δ Sct and eclipsing binary).
- Inclusion of the full variability catalog on both TASOC and MAST as an new high-level data product.

¹⁸ <https://archive.stsci.edu/hlsp/stella>

¹⁹ <https://tasoc.dk>

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement N°670519: MAMSIE), from the KU Leuven Research Council (grant C16/18/005: PARADISE), from the Research Foundation Flanders (FWO) under grant agreement G0H5416N (ERC Runner Up Project), as well as from the BELgian federal Science Policy Office (BELSPO) through PRODEX grant PLATO. D.J.A acknowledges support from the STFC via an Ernest Rutherford Fellowship (ST/R00384X/1). Funding for the Stellar Astrophysics Centre is provided by The Danish National Research Foundation (Grant agreement no.: DNRFF106). RH and MNL acknowledges the ESA PRODEX programme. This research was supported by the National Aeronautics and Space Administration (80NSSC18K1585 and 80NSSC19K0379) awarded through the TESS Guest Investigator Program. K.J.B. is supported by the National Science Foundation under Award AST-1903828. J.S.K and K.J.B. were supported by funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no. 338251 (StellarAges). DMB gratefully acknowledges funding from a senior postdoctoral fellowship from the Research Foundation Flanders (FWO) with grant agreement No. 1286521N. The research leading to these results has received funding from the Research Foundation Flanders (FWO) under grant agreement G0A2917N (BlackGEM). R.A.G. acknowledges the support from the GOLF and PLATO CNES grants. L.M. and E.M. were supported by the Premium Postdoctoral Research Program of the Hungarian Academy of Sciences. The research leading to these results has been supported by the Hungarian National Research, Development and Innovation Office (NKFIH) grant KH_18 130405 and the Lendület LP2014-17 and LP2018-7/2020 grants of the Hungarian Academy of Sciences. D.B. acknowledges support from the NASA TESS Guest Investigator Program under award 80NSSC19K0385.

This paper includes data collected by the TESS mission, which are publicly available from the Mikulski

Archive for Space Telescopes (MAST) and described in Jenkins et al. (2016). Funding for the TESS mission is provided by NASA's Science Mission Directorate. This research has made use of NASA's Astrophysics Data System, as well as the NASA/IPAC Extragalactic Database (NED) which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration. Funding for the TESS Asteroseismic Science Operations Centre is provided by the Danish National Research Foundation (Grant agreement no.: DNRFF106), ESA PRODEX (PEA 4000119301) and Stellar Astrophysics Centre (SAC) at Aarhus University. We thank the TESS team and staff and TASC/TASOC for their support of the present work.

This paper includes data collected by the Kepler mission. Funding for the Kepler and K2 mission was provided by NASA's Science Mission Directorate. The authors acknowledge the efforts of the Kepler Mission team in obtaining the light curve data and data validation products used in this publication. These data were generated by the Kepler Mission science pipeline through the efforts of the Kepler Science Operations Center and Science Office. The Kepler light curves are archived at the Mikulski Archive for Space Telescopes.

The numerical results presented in this work were obtained at the Centre for Scientific Computing, Aarhus²⁰. This research made use of Astropy, a community-developed core Python package for Astronomy (Astropy Collaboration et al. 2013, 2018).

Software: Scikit-learn (Pedregosa et al. 2011), Numpy (Harris et al. 2020), Astropy (Astropy Collaboration et al. 2013, 2018), Scipy (Virtanen et al. 2020), Pandas (Wes McKinney 2010; pandas development team 2020), Lightkurve (Lightkurve Collaboration et al. 2018), XGBoost (Chen & Guestrin 2016), Tensorflow (Abadi et al. 2015)

REFERENCES

- Abadi, M., Agarwal, A., Barham, P., et al. 2015, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, software available from tensorflow.org

²⁰ <https://phys.au.dk/forskning/cscaa/>

- Abdul-Masih, M., Prša, A., Conroy, K., et al. 2016, *AJ*, 151, 101
- Aerts, C., Christensen-Dalsgaard, J., & Kurtz, D. W. 2010, *Asteroseismology*
- Aerts, C., Eyer, L., & Kestens, E. 1998, *A&A*, 337, 790
- Aggarwal, C. C. 2014, *Data Classification: Algorithms and Applications*, 1st edn. (Chapman & Hall/CRC)
- Antoci, V., Cunha, M. S., Bowman, D. M., et al. 2019, *MNRAS*, 490, 4040
- Armstrong, D. J., Kirk, J., Lam, K. W. F., et al. 2016, *Monthly Notices of the Royal Astronomical Society*, 456, 2260
- Armstrong, D. J., Kirk, J., Lam, K. W. F., et al. 2015, *Astronomy and Astrophysics*, 579, A19
- Astropy Collaboration, Price-Whelan, A. M., Sipőcz, B. M., et al. 2018, *AJ*, 156, 123
- Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, *A&A*, 558, A33
- Auvergne, M., Bodin, P., Boissard, L., et al. 2009, *A&A*, 506, 411
- Bae, J., Ryu, Y., Chang, T., Song, I., & Kim, H. M. 1996, *Signal Processing*, 52, 75
- Ball, N. M., Brunner, R. J., Myers, A. D., & Tcheng, D. 2006, *ApJ*, 650, 497
- Borucki, W. J., Koch, D., Basri, G., et al. 2010, *Science*, 327, 977
- Bowman, D. M. 2020, *Frontiers in Astronomy and Space Sciences*, 7, 70
- Bowman, D. M., Kurtz, D. W., Breger, M., Murphy, S. J., & Holdsworth, D. L. 2016, *MNRAS*, 460, 1970
- Bracewell, R. N. 1986, *The Fourier Transform and its applications*
- Breiman, L. 2001, *Machine Learning*, 45, 5
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. 1984, *Classification and regression trees*, The Wadsworth statistics/probability series (Belmont: Wadsworth)
- Brett, D. R., West, R. G., & Wheatley, P. J. 2004, *Monthly Notices of the Royal Astronomical Society*, 353, 369
- Brown, T. M., Gilliland, R. L., Noyes, R. W., & Ramsey, L. W. 1991, *ApJ*, 368, 599
- Bruntt, H. & Buzasi, D. L. 2006, *Mem. Soc. Astron. Italiana*, 77, 278
- Bugnet, L., García, R. A., Davies, G. R., et al. 2018, *A&A*, 620, A38
- Busa, M. A. & van Emmerik, R. 2016, *Journal of Sport and Health Science*, 5, 44
- Buzasi, D. L. 2004, in *ESA Special Publication*, Vol. 538, *Stellar Structure and Habitable Planet Finding*, ed. F. Favata, S. Aigrain, & A. Wilson, 205–213
- Chen, T. & Guestrin, C. 2016, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16 (New York, NY, USA: ACM), 785–794
- Costa, M., Goldberger, A. L., & Peng, C.-K. 2005, *Phys. Rev. E*, 71, 021906
- Debusscher, J., Blomme, J., Aerts, C., & De Ridder, J. 2011, *A&A*, 529, A89
- Debusscher, J., Sarro, L. M., Aerts, C., et al. 2007, *A&A*, 475, 1159
- Degroote, P., Aerts, C., Ollivier, M., et al. 2009, *A&A*, 506, 471
- Derekas, A., Plachy, E., Molnár, L., et al. 2017, *MNRAS*, 464, 1553
- Eyer, L. & Blake, C. 2005, *MNRAS*, 358, 30
- Eyer, L. & Grenon, M. 1998, in *New Eyes to See Inside the Sun and Stars*, ed. F.-L. Deubner, J. Christensen-Dalsgaard, & D. Kurtz, Vol. 185, 291
- Fawcett, T. 2006, *Pattern Recognition Letters*, 27, 861, *rOC Analysis in Pattern Recognition*
- Feinstein, A. D., Montet, B. T., Ansdell, M., et al. 2020, *AJ*, 160, 219
- Friedman, J. H. 2001, *The Annals of Statistics*, 29, 1189
- Gaia Collaboration, Eyer, L., Rimoldini, L., et al. 2019, *A&A*, 623, A110
- Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., et al. 2016, *A&A*, 595, A1
- García, R. A. & Ballot, J. 2019, *Living Reviews in Solar Physics*, 16, 4
- García, R. A., Ceillier, T., Salabert, D., et al. 2014, *A&A*, 572, A34
- Granitto, P. M., Furlanello, C., Biasioli, F., & Gasperi, F. 2006, *Chemom Intell Lab Syst*, 83, 83
- Guzik, J. A., Kaye, A. B., Bradley, P. A., Cox, A. N., & Neuforge, C. 2000, *ApJL*, 542, L57
- Hall, M. A. 1999, PhD thesis, University of Waikato Hamilton
- Handberg, R., Lund, M. N., White, T. R., et al. 2021, *arXiv e-prints*, arXiv:2106.08341
- Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, *Nature*, 585, 357
- Hekker, S. & Christensen-Dalsgaard, J. 2017, *A&A Rv*, 25, 1
- Hon, M., Stello, D., & Zinn, J. C. 2018, *The Astrophysical Journal*, 859, 64
- Howell, S. B., Sobeck, C., Haas, M., et al. 2014, *PASP*, 126, 398
- Hümmerich, S., Mikulášek, Z., Paunzen, E., et al. 2018, *A&A*, 619, A98

- Jamal, S. & Bloom, J. S. 2020, arXiv e-prints, arXiv:2003.08618
- Kallinger, T. 2019, arXiv e-prints, arXiv:1906.09428
- Kedem, B. & Slud, E. 1981, *Biometrika*, 68, 551
- Kedem, B. & Slud, E. 1982, *The Annals of Statistics*, 10, 786
- Kgoadi, R., Engelbrecht, C., Whittingham, I., & Tkachenko, A. 2019, arXiv preprint arXiv:1906.06628
- Kim, D.-W. & Bailer-Jones, C. A. L. 2016, *Astron Astrophys*, 587, A18
- Kirk, B., Conroy, K., Prša, A., et al. 2016, *AJ*, 151, 68
- Kiss, L. L. & Bódi, A. 2017, *A&A*, 608, A99
- Kjeldsen, H. & Bedding, T. R. 1995, *A&A*, 293, 87
- Kjeldsen, H., Bedding, T. R., Viskum, M., & Frandsen, S. 1995, *AJ*, 109, 1313
- Kohonen, T. 1990, *Proceedings of the IEEE*, 78, 1464
- Kozachenko, L. F. & Leonenko, N. N. 1987, *Probl. Peredachi Inf.*, 23, 95–101
- Kraskov, A., Stögbauer, H., & Grassberger, P. 2004, *Phys. Rev. E*, 69, 066138
- Kuslewicz, J. S., Hekker, S., & Bell, K. J. 2020, *MNRAS*, 497, 4843
- Li, G., Van Reeth, T., Bedding, T. R., et al. 2020, *MNRAS*, 491, 3586
- Lightkurve Collaboration, Cardoso, J. V. d. M., Hedges, C., et al. 2018, *Lightkurve: Kepler and TESS time series analysis in Python*, *Astrophysics Source Code Library*
- Lomb, N. R. 1976, *Ap&SS*, 39, 447
- Lundberg, S. M., Erion, G., Chen, H., et al. 2020, *Nature Machine Intelligence*, 2, 2522
- Lundberg, S. M. & Lee, S.-I. 2017, in *Advances in Neural Information Processing Systems 30*, ed. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Curran Associates, Inc.), 4765–4774
- Manick, R., Kamath, D., Van Winckel, H., et al. 2019, *A&A*, 628, A40
- Mathys, G., Kurtz, D. W., & Holdsworth, D. L. 2020, *A&A*, 639, A31
- Matijević, G., Prša, A., Orosz, J. A., et al. 2012, *AJ*, 143, 123
- McQuillan, A., Mazeh, T., & Aigrain, S. 2014, *ApJS*, 211, 24
- Modak, S., Chattopadhyay, T., & Chattopadhyay, A. K. 2018, arXiv e-prints, arXiv:1801.09406
- Montgomery, M. H. & Odonoghue, D. 1999, *Delta Scuti Star Newsletter*, 13, 28
- Namekata, K., Maehara, H., Notsu, Y., et al. 2019, *ApJ*, 871, 187
- Naul, B., Bloom, J. S., Pérez, F., & van der Walt, S. 2018, *Nature Astronomy*, 2, 151
- Nielsen, M. B., Gizon, L., Schunker, H., & Karoff, C. 2013, *A&A*, 557, L10
- pandas development team, T. 2020, pandas-dev/pandas: Pandas
- Pápics, P. I., Briquet, M., Baglin, A., et al. 2012, *A&A*, 542, A55
- Pápics, P. I., Tkachenko, A., Van Reeth, T., et al. 2017, *A&A*, 598, A74
- Pedersen, M. G., Escorza, A., Pápics, P. I., & Aerts, C. 2020, *MNRAS*, 495, 2738
- Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, *Journal of Machine Learning Research*, 12, 2825
- Pinsonneault, M. H., Elsworth, Y. P., Tayar, J., et al. 2018, *ApJS*, 239, 32
- Plachy, E., Bódi, A., & Kolláth, Z. 2018, *MNRAS*, 481, 2986
- Pojmanski, G. 2002, *AcA*, 52, 397
- Preston, G. W. 1974, *ARA&A*, 12, 257
- Provost, F. 2000, in *Proceedings of the AAAI'2000 workshop on imbalanced data sets* (AAAI Press)
- Prša, A., Batalha, N., Slawson, R. W., et al. 2011, *AJ*, 141, 83
- Rauer, H., Catala, C., Aerts, C., et al. 2014, *Experimental Astronomy*, 38, 249
- Richards, J. W., Starr, D. L., Butler, N. R., et al. 2011, *Astrophys J*, 733 [1101.1959]
- Richman, J. S. & Moorman, J. R. 2000, *American Journal of Physiology-Heart and Circulatory Physiology*, 278, H2039, PMID: 10843903
- Ricker, G. R., Winn, J. N., Vanderspek, R., et al. 2015, *Journal of Astronomical Telescopes, Instruments, and Systems*, 1, 014003
- Roberts, D. H., Lehar, J., & Dreher, J. W. 1987, *AJ*, 93, 968
- Santos, A. R. G., García, R. A., Mathur, S., et al. 2019, *ApJS*, 244, 21
- Sarro, L. M., Debosscher, J., López, M., & Aerts, C. 2009, *A&A*, 494, 739
- Scargle, J. D. 1982, *ApJ*, 263, 835
- Schapire, R. E. 1990, *Machine Learning*, 5, 197
- Shannon, C. E. 1948, *Bell System Technical Journal*, 27, 623
- Shapiro, S. S. & Wilk, M. B. 1965, *Biometrika*, 52, 591
- Slawson, R. W., Prša, A., Welsh, W. F., et al. 2011, *AJ*, 142, 160
- Stassun, K. G., Oelkers, R. J., Paegert, M., et al. 2019, *AJ*, 158, 138

- Szabó, R. 2018, in *The RR Lyrae 2017 Conference. Revival of the Classical Pulsators: from Galactic Structure to Stellar Interior Diagnostics*, Vol. 6, 119–123
- Szabó, R., Szabados, L., Ngeow, C. C., et al. 2011, *MNRAS*, 413, 2709
- Thompson, S. E., Caldwell, D. A., Jenkins, J. M., et al. 2016, *Kepler Data Release 25 Notes*, Kepler Science Document KSCI-19065-002
- Tkachenko, A., Aerts, C., Yakushechkin, A., et al. 2013, *A&A*, 556, A52
- Udalski, A., Szymański, M. K., Soszyński, I., & Poleski, R. 2008, *AcA*, 58, 69
- Udalski, A., Szymanski, M. K., & Szymanski, G. 2015, *AcA*, 65, 1
- Valenzuela, L. & Pichara, K. 2018, *MNRAS*, 474, 3259
- Van Reeth, T., Tkachenko, A., & Aerts, C. 2016, *A&A*, 593, A120
- Van Reeth, T., Tkachenko, A., Aerts, C., et al. 2015a, *A&A*, 574, A17
- Van Reeth, T., Tkachenko, A., Aerts, C., et al. 2015b, *ApJS*, 218, 27
- Vega, L. D., Stassun, K. G., Montez, Rodolfo, J., Boyd, P. T., & Somers, G. 2017, *ApJ*, 839, 48
- Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, *Nature Methods*, 17, 261
- Waelkens, C., Aerts, C., Kestens, E., Grenon, M., & Eyer, L. 1998, *A&A*, 330, 215
- Walker, G., Matthews, J., Kuschnig, R., et al. 2003, *PASP*, 115, 1023
- Wes McKinney. 2010, in *Proceedings of the 9th Python in Science Conference*, ed. Stéfan van der Walt & Jarrod Millman, 56 – 61
- Wolpert, D. H. 1992, *Neural Networks*, 5, 241
- Wyrzykowski, L. & Belokurov, V. 2008, in *American Institute of Physics Conference Series*, Vol. 1082, *Classification and Discovery in Large Astronomical Surveys*, ed. C. A. L. Bailer-Jones, 201–206
- Youden, W. J. 1950, *Cancer*, 3, 32
- Yu, J., Bedding, T. R., Stello, D., et al. 2020, *MNRAS*, 493, 1388

APPENDIX

A. FEATURE IMPORTANCE PLOTS

The SHAP²¹ feature importance plots (see Sect 6.2 for an explanation) allow us to evaluate the importance of the different attributes used by each classifier on a per class basis. The hatched regions in the plots indicate the most important feature per class. The plots for RFGC, SORTING-HAT and GBGC are respectively shown in Figs. 12, 13 and 14. Due to the fact that updated features have been used in the training of RFGC, the feature importances are different to those reported in [Armstrong et al. \(2016\)](#).

For RFGC this reveals that the zero-crossings parameter is the most important for classifying contactEB/spots and dSct/bCep stars. The point-to-point differences are the primary features for solar-like oscillators and aperiodic stars, while respectively the coherency parameter, first fundamental period, FliPer value and the SOM are the most important for constant, gDor/SPB, RRLyr/Ceph and transit/eclipse stars.

In the case of SORTING-HAT we notice that the multiscale Entropy is by far the most important, as it is the primary feature for the contactEB/spots, aperiodic, constant, solar and gDor/SPB classes. In addition to that, the differential entropy is the primary feature for RRLyr/Ceph stars. For transit/eclipse stars the skewness is the most important followed by the flux ratio, which is logical given that these types of stars tend to have very skewed light curves. Lastly, for dSct/bCep stars the first fundamental frequency is most important.

For GBGC the variability index is the primary feature for constant, aperiodic and dSct/bCep stars. The range of the cumulative sum of the fluxes of the phase-folded light curve is the primary feature for solar-like oscillators and contactEB/spots stars. The skewness, first fundamental period and Shapiro-Wilk test for normality of the light curve are respectively the primary features for the transit/eclipse, gDor/SPB and RRLyr/Ceph classes.

B. EFFECTS OF PHOTON NOISE

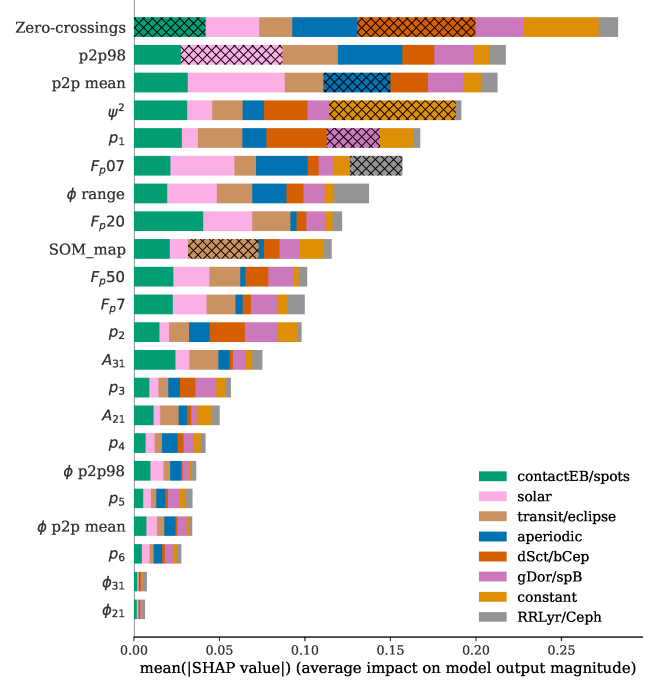
C. CLASS PROBABILITIES *KEPLER* Q9

Figure 12. RFGC feature importances from SHAP. The hatched regions indicate the most important feature per class.

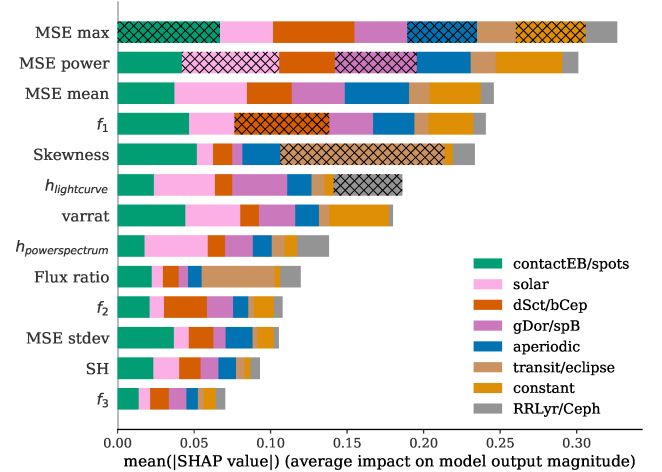


Figure 13. SORTING-HAT feature importances from SHAP. The hatched regions indicate the most important features per class.

²¹ <https://github.com/slundberg/shap>

Table 6. First five rows of the electronic table with class probabilities and assigned labels.

kie	p(aperiodic)	p(constant)	p(contactEB/spots)	p(dSct/bCep)	p(eclipse/transit)	p(gDor/SPB)	p(RRLyr/Ceph)	p(solar)	Label	Label (Youden)
11513597	0.0641	0.0221	0.5526	0.0354	0.0327	0.1211	0.1078	0.0642	contactEB/spots	contactEB/spots
9266835	0.0299	0.0121	0.5285	0.0639	0.0819	0.1530	0.1001	0.0305	contactEB/spots	contactEB/spots
11241837	0.02449	0.0012	0.7854	0.01607	0.0259	0.0477	0.0858	0.0134	contactEB/spots	contactEB/spots
9591826	0.0362	0.0849	0.3630	0.1165	0.0542	0.1967	0.0971	0.0514	contactEB/spots	contactEB/spots

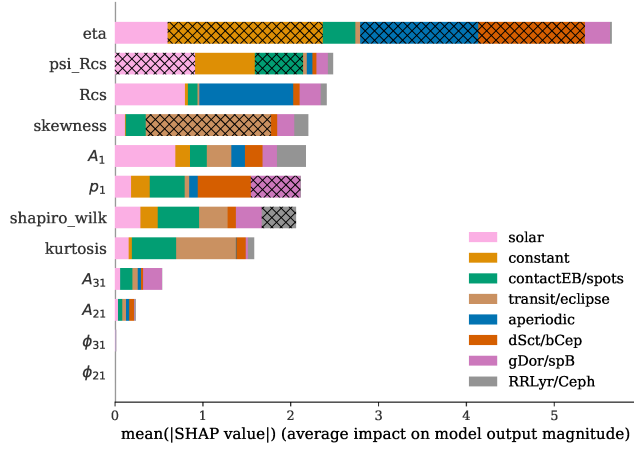


Figure 14. GBGC feature importances from SHAP. The hatched regions indicate the most important feature per class.

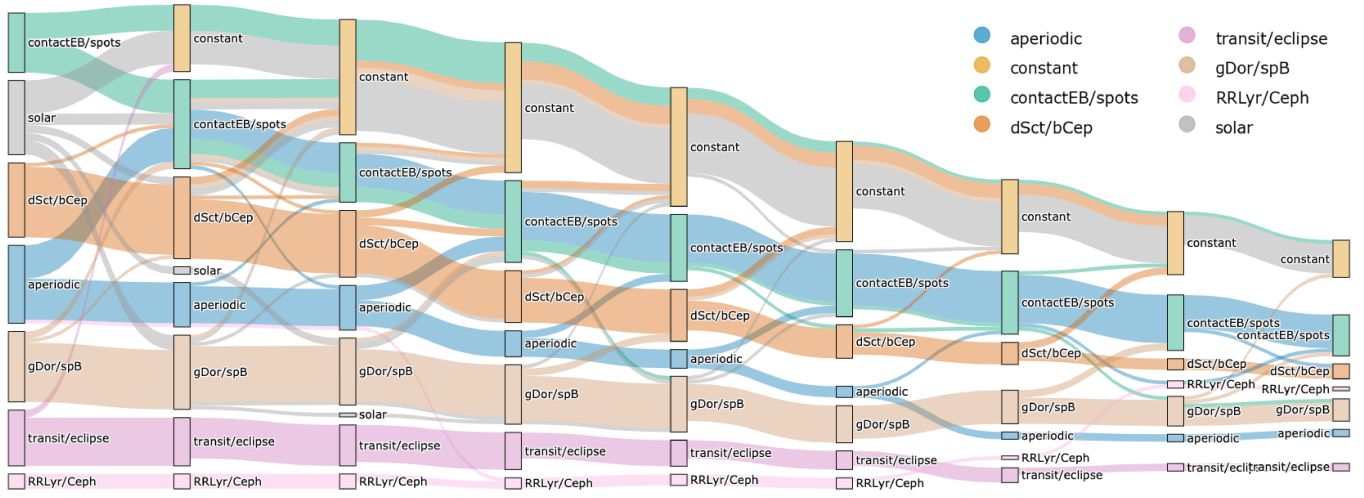


Figure 15. Sankey plot indicative of how stars move between classes when more noise is added to their light curves. The relatively brightest stars are shown on the left and relatively faintest on the right. The colors of the streams indicate the true label of the stars therein, and the bar labels and colors are representative of the predicted class. The height of the bars corresponds to the number of stars in that bin. The number of stars decreases from left to right because stars are eliminated once their newly calculated magnitude exceeds 15.5. Each step corresponds to a noise increase representative of 0.5 magnitude. Step 0 shows the original predictions by the metaclassifier.

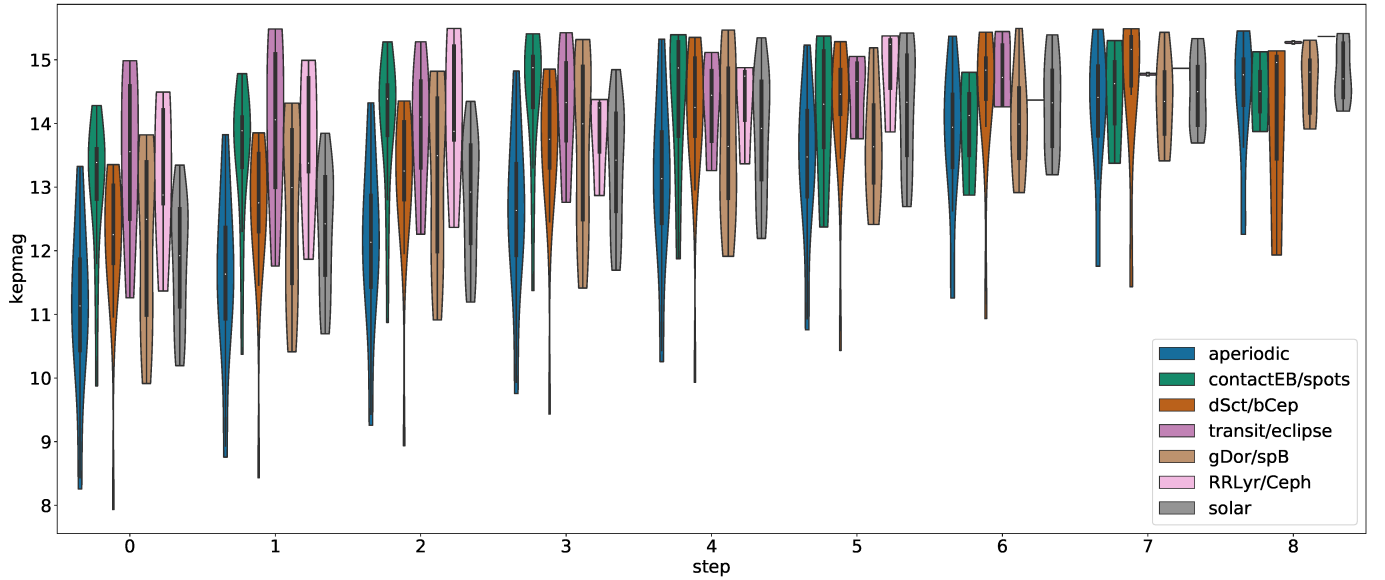


Figure 16. Violin plot illustrating the magnitude distribution at each step or bar of the Sankey plot in Fig. 15.

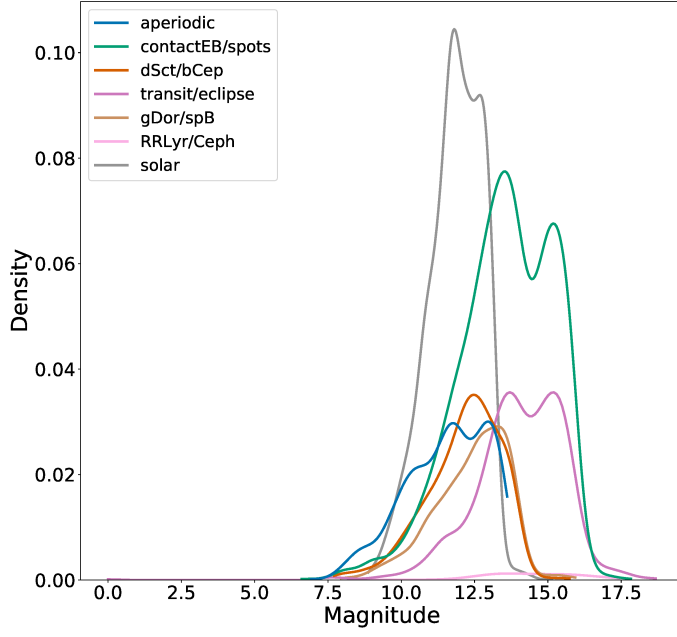


Figure 17. KDE plot illustrating the magnitude distribution of all stars in the training set described in Sect. 3.