Machine Learning for Variable Stars in K2 Data – Interim Report

Chance Haycock

Department of Physics, University of Warwick, Coventry CV4 7AL, United Kingdom

I. Introduction

Due to the extensive amount of data collected by today's astronomical surveys, the need for automated classification of celestial bodies is on the increase. Modern data analysis techniques, which fall under the umbrella term of *Machine Learning*, are consequently becoming an increasingly useful tool in the field of astronomy; managing to digest the vast amounts of data efficiently and hence, further develop our understanding of the wider universe. *Variable stars*, defined as stars whose brightness changes over time, have been of interest to astronomers for many years and so in more recent times, the classification of these stars into their associated classes has become a suitable problem for machine learning to tackle[1, 2]. Two of the most useful types of variable stars are *RR Lyrae* and *Eclipsing Binaries*; both of which have distinctive *lightcurves* (See Section C..1) which are central to the studies in this paper. In particular, RR Lyrae are of great use to astronomers as standard distance candles[3], with new discoveries improving the precision to which they are measured. Eclipsing binaries are also useful for estimating distances to neighbouring galaxies and consequently, increasingly accurate estimates of the Hubble constant[4, 5]. Although there are possibly several causes of variability (See Section II.A.), a recent study of massive stars observed low frequency variability and concluded its cause to be due to internal gravity waves[6], leading to the inference of important constraints on the structure of massive stars and their evolution[7]. Studies like this demonstrate the relevance of variable stars in today's astrophysics community and the advantages of developing an automated classification model. Observing changes in brightness of distant stars is also being used to search for objects other than variable stars. Large scale surveys such as NASA's Transiting Exoplanet Survey[8] (TESS) are primarily searching

for exoplanets, possibly habitable, by observing changes in a host star's apparent brightness known as *transits*. Observations are made for roughly 200,000 of the brightest stars near Earth spanning a total mission duration of two years. However, in this work we study data collected from another of NASA's recent missions; K2 - the revived Kepler mission[9]. The mission is split into 20 campaigns of roughly 20,000 targets each observed for approximately 80 days at 30 minute cadences. As of December 2019, raw data is available for download[1] for campaigns 0-19. Another, perhaps more ambitious, ongoing mission which will be of use to us here is the European Space Agency's GAIA[10] mission; attempting to map the whole sky, producing the largest and most accurate space catalogue to date. Using data from the second GAIA data release[11] along with lightcurve data from the K2 mission, we aim here to produce a *catalogue* of roughly 200,000 variable stars complete with classification probabilities for K2 campaigns 5 and upwards using machine learning techniques; extending work as carried out by *Armstrong et. al. 2016*[12]. In essence, our final model will learn from the classifications made in campaigns 0-4 by Armstrong et. al. to make predictions of the classes that targets in campaigns 5-19 belong to.

[1] https://archive.stsci.edu/k2/

II. Theory

A. Variable Star Classes

For the work carried out here, we restrict ourselves to classifying the data into the same 7 distinct classes used in Armstrong et. al. 2016; namely RR Lyraes, detached Eclipsing Binaries, semi-detached/contact Eclipsing Binaries, γ-Dors, δ-Scuti, other periodic variables and Noise. Most of these variable types are located on the instability strip[13] of the famous Hertzsprung-Russell (HR) diagram. Most stars that lie on the instability strip, particularly classic Cepheids, are varying in brightness due to partial ionisation of helium and hydrogen; a process more commonly referred to as the κ-mechanism. The distinction between these variable classes is often based on the star's period, the cause of brightness variation and the order of magnitude on which the brightness varies.

A..1 δ-Scuti (DSCUT)

Sitting closest to the strip of main sequence stars on the HR diagram, δ-Scuti variables are varying due to both radial *and* non-radial pulsations of the star. The former can be attributed to the aforementioned κ-mechanism, whilst the latter is due to parts of the surface moving in opposite directions. Typical changes in magnitude are in the range of 0.1-1.0 magnitudes with a

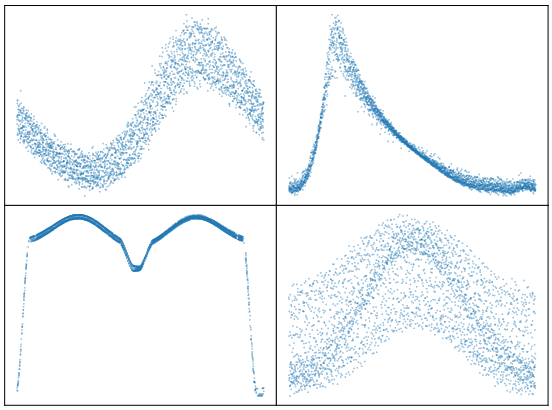

Figure 1: Starting from top left and reading off clockwise, we present typical phase-folded lightcurves for DSCUT, RRab, GDOR and EB variable stars corresponding to EPIC IDs 206032188, 206397568, 206382857, 206202136 respectively from campaign 3 of the K2 mission as classified by Armstrong et. al.

typical period of a few hours[14]. The adherence to a period-luminosity relation also makes these variables useful distance estimators, helping astronomers estimate distances to the galactic centre[15]. The typical light curves of δ-Scutis can be characterised by a sinusoidal variation of multiple frequency modes as shown in Figure 1.

A..2 Eclipsing Binaries (EA & EB)

As the name suggests, Eclipsing Binaries ares systems of *two* stars orbiting their common centre of mass. Both types of Eclipsing Binaries are *extrinsic* variable stars; meaning both stars do not actually vary in brightness. The observed variation is due to one of the star's orbiting companion periodically eclipsing it. The light curves of both types can be uniquely characterised by their primary and secondary eclipses, coinciding with the largest and smallest dip in apparent brightness respectively. EA stars (also known as Algol type) are typically characterised by well defined beginnings of primary and secondary eclipses separated by periods of constant brightness. EB stars orbit each other closer than EA systems and so the distinction between the beginning of eclipses is harder to infer. Both types have a typical period of variability between 0.5-2 days and a typical change in brightness of 1 magnitude.

A..3 γ-Dors (GDOR)

γ-Dor variables are a relatively new class of variable star, sitting towards the red end of the δ-Scuti instability strip. Often compared with δ-Scutis, γ-Dor variables tend to have slight longer periods in the range of 0.5-3 days and variations in brightness driven by non-radial pulsations in a process known as convective blocking[16]. Some literature refers to this as pulsating in high order g-modes. Their lightcurves also exhibit sinusoidal behaviour with typical magnitude fluctuations of the order of 0.1 magnitudes. Theory predicts a region of overlap between instability strips on the HR diagram in which hybrids of both types of star should exist. Results from the Kepler mission show that nearly all δ-Scuti stars exhibit γ-Dor-like behaviour reducing the purity of these two particular classes.

A..4 RR Lyrae (RRab)

In general, RR Lyraes exist in 3 different flavours, RRab, RRc, RRd, with 91% of all observed RR Lyraes falling into the RRab category[17]; pulsating in their fundamental mode. In particular, RRab's possess distinctive asymmetric lightcurves featuring steep rises in brightness, which along with a well obeyed period-luminosity relation in the infrared window of the electromagnetic spectrum make them a great tool for astronomers[18]. Period of variability is typically in the range of 4 hours - 1 day and the radial pulsation is again driven by the κ-mechanism with changes in brightness of the order of 1 magnitude. i.e. RR Lyraes are similar to classic Cepheids but with shorter periods and lower luminosity.

B. K2 - Kepler's Second Light

Launched in 2009, NASA's original Kepler mission came to a sudden halt after the second of four reaction wheels had broken; consequently the spacecraft was not able to maintain a consistent field of view. In its 4 years of transmission, Kepler is said to have revolutionised the study of exoplanets and astrophysics[9], achieving photometric precision as low as 15 ppm, leading to the discovery of exoplanets such as Kepler-4b within its first 6 weeks[19]. Fortunately, engineers were able to reinvent the spacecraft by using solar pressure and periodic thrusting? to maintain the pointing of the spacecraft; reinventing the original Kepler as *K2*. The new setup requires the spaceship to point at a new field of view roughly every three months; dividing the mission into what we here on refer to as *campaigns* displayed in Figure 2. However, as expected, the K2 raw aperture photometry is between 3-4 times less precise than Kepler and an instrumental noise is additionally introduced[20]. Several methods have therefore been introduced to remedy this problem with some of the most successful achieving half the precision of Kepler.

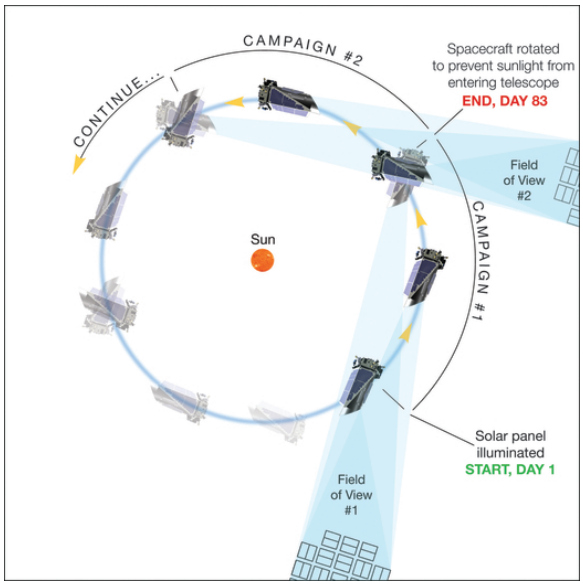

Figure 2: The plan of the K2 mission. Data is available for 20 campaigns, each observed for roughly 80 days at 30 minute cadences. Figure taken from Howell et. al. [9].

C. Sources of Data

Several authors have recently provided intricate methods for reducing the instrumental noise present in the K2 lightcurves; some of the most popular include the `K2SC`, `K2SFF`, `K2VARCAT` (developed and used in Armstrong et. al. (2016)) and `EVEREST` pipelines; all available as high level science products (HLSP) on the MAST website[2]. These pipelines behave in different ways depending on the observed brightness of the target; some achieving higher precision when compared to the original Kepler mission and others being better suited to variability studies. We discuss this particular trade-off in Section C..1?. The work carried out in this paper primarily makes use of the K2SC pipeline which has available processed data for K2 campaigns 3-8, 10 and is highly recommended for variability studies[21].

C..1 K2 Systematics Correction

The K2SC pipeline tackles the precision problem by using Gaussian processes (GP) to model *both* the instrumental systematics and astrophysical variability of targets independently, giving the user the option to remove either or both. As we focus our studies on the variability of targets, the work carried out here removes only the position-dependent instrumental noise, preserving the time-dependent variability of targets. Users searching for exoplanet transits are advised to remove both sources of noise. The pipeline handles lightcurves with two different kernels; quasi-periodic and non-periodic. By computing the Lomb-Scargle periodogram[22] of the light curve and comparing the false-alarm probability of the outputted period with a pre-determined value, the pipeline is able to choose the appropriate regime. For comparison, the time component of the Gaussian process covariance function in the non-periodic regime is described by

$$k_t(t_i, t_j) = A_t \exp[-\eta_t (t_i - t_j)^2]$$

where A_t is the amplitude and η_t the inverse length scale. In the quasi-periodic regime, it is replaced with

$$k_t(t_i, t_j) = A_t \exp\left[-\Gamma \sin^2\left(\frac{\pi |t_i - t_j|}{P}\right) - \frac{(t_i - t_j)^2}{L_e^2}\right]$$

where P is the period, Γ the inverse length scale of the periodic component of the variations, and L_e the evolutionary timescale of the variations[21]. In both instances, the input parameters which maximise the standard likelihood function are used in the model. Measured with a 6-hour CDPP[3], the K2SC pipeline achieves a precision within a factor of 1.5 of the origi-

[2] https://archive.stsci.edu/k2/hlsps.html

[3] Combined Differential Photometric Precision - Formally defined as the inverse signal-to-noise ratio of a reference signal of the corresponding duration, in parts per million (ppm)

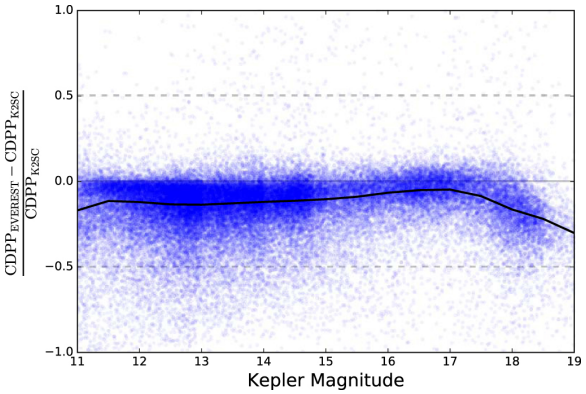

Figure 3: Relative 6-hour CDPP for EVEREST vs. K2SC for each lightcurve in campaigns 0-7 taken from Luger et. al. (2016). The black line shows a median taken over bin widths of 0.5 magnitudes. Points which fall under the zero line are lightcurves which achieve higher precision with the EVEREST pipeline.

nal Kepler mission for bright stars observed with a Kepler magnitude less than 12. Consequently, the K2SC lightcurves have been found to be considerably less noisy than the K2VARCAT lightcurves and similar, if not more precise, to K2SFF lightcurves. We opt to use the K2SC lightcurves here over K2SFF due to K2SC's robustness with astrophysical variability; the subject of our study. However, more recently the EVEREST pipeline has been found to obtain higher precision than all previously mentioned pipelines for Kepler magnitudes $>$ 11; achieving the *same* precision as the original Kepler mission for bright stars with magnitudes $<$ 13. Due to this success, the EVEREST pipeline may be used in future work to compare results obtained with the K2SC pipeline; with the additional advantage of available data for every K2 campaign.

D. Machine Learning Techniques

Machine learning techniques fall under two branches, namely *supervised* and *unsupervised*; both of which can be used for classification problems. The distinction between the two is whether they have prior knowledge of the underlying true label or class of a data point. Therefore, the goal of an unsupervised algorithm is to group data points based on their natural structure whilst the goal of a supervised algorithm is to learn a function which best fits data points to their known classes. In this work and Armstrong et. al. (2016), an unsupervised algorithm known as a self-organising map (SOM), and a supervised algorithm known as a Random Forest (RF) are used in conjunction with each other to obtain classifications of the observed K2 targets. The overall methodology of the automated classification process is to represent each K2 target by a finite number of data *features* which are chosen/calculated by the user. Using a *training set* (a subset of *known* data types), these features along with the known class for each light curve are inputted into the RF as a finite list of real numbers. By minimising a *loss function*, which corresponds to the model mislabelling known types, the best fitting model can be found and then used to classify unseen data. Alternatively, another subset of known data referred to as a *test set* can be

passed into the existing RF model to quantify how well the model classifies known data types. Tweaking parameters of the RF model and different choices of training and test sets can influence the performance of the model and so it is the user's task to tune appropriately to increase the model's performance.

D..1 SOM/Kohonen Maps

A SOM is typically used to extract information about the shape of an object or image; in this case, the shape of the phase-folded lightcurves similar to the ones shown in Figure 1. In order to pass these shapes into the SOM, the phase-folded lightcurves are typically binned by their mean value in N bins of equal width. The SOM then interprets the shape as a N-dimensional array or real numbers and groups other objects together according to similarity by the use of a *Kohonen Layer*. The Kohonen layer consists of pixels which each represent a template to compare the inputted lightcurve against. The best matching template (typically measured by smallest L^2 distance) then represents the best matching pixel on the map. For visualisation, it is typically best to reduce the N-dimensional object by using a 2D Kohonen layer as shown in Figure 5 and producing a 2D map of data points as shown in Figure 4. The co-ordinates outputted by the SOM for a given lightcurve can be used as a feature in the RF, inferring properties about the shape of the variability of the star.

D..2 Random Forests

Aside from the shape of the lightcurve, there are several other features that may help to distinguish one class from another in the RF model such as period of variability, absolute magnitude, magnitude of variability etc. The RF, made up of a large number of small estimators known as decision trees, handles this classification task by using two sources of randomness. Each decision tree in the RF is built from a subsample of the training data with replacement. Additionally, each tree makes decisions (splitting of nodes) based on a random sub-samples of the input features; typically of size approximately the square root of the full data feature set. The purpose of this is to reduce the variance of the overall RF yielding an overall better model[23]. Continuing on from the success of the use of a RF to classify variable stars (see Richards et. al. 2011[24]), we initially use a RF model in this work, perhaps exploring other classifiers in the future.

III. Work Done in Term 1

Term 1 has been used to extend the work as carried out in Armstrong et. al. 2016 to later K2 campaigns using similar techniques and data features. The aim was to have a working model which correctly classifies known stars from campaigns 3 and 4.

A. Data Collection

In contrast to the work aforementioned, we have opted to use lightcurves which have been pre-processed as in Section C..1 to calculate features of the lightcurve such as period, point-to-point (p2p from here on) scatter and skew. In addition to the instrumental systematics correction, we have additionally reduced the effects of long term (> 20 days) variability by fitting a 3rd degree polynomial to the $\approx$ 80 day K2SC lightcurve. By doing this, we obtain cleaner phase-folds of the lightcurve and an implicit normalisation of p2p statistics. Additionally, from the 29,600 targets in K2 campaigns 3 and 4, we have cross-matched data for 27,525 of those targets from the GAIA mission. In particular, this data includes an estimate of the distance to the stars calculated by [25].

B. Automated Period Finding

From previous work, it is apparent that the period of variability of the star is an important feature for classification. To do this, we have used the reliable Lomb-Scargle algorithm to extract the two most dominant periods of the lightcurve and the ratio of their power amplitudes[4].

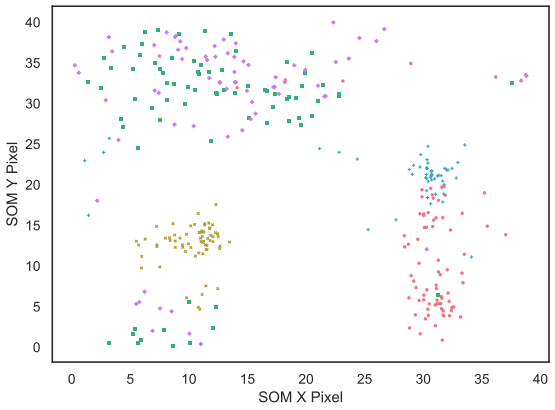

Figure 4: A visualisation of the training set for the SOM. We manage to obtain clustering for RRab's, EA's and EB's but the SOM fails to distinguish between DSCUT's and GDOR's well as expected. Note: Random jitter has been added to each pixel for distinction.

C. Training Sets

For both the SOM and Random Forest, a training set is required. Initially, we used the entire data from campaigns 3 and 4 but discovered that some classifications were made with probabilities as low as 0.3. To overcome this, the training set was restricted to sufficient probability thresholds to ensure that training sets comprised of clean and very likely classifications. Figure 4 shows how this data clusters on a SOM and Figure 5 shows the final Kohonen layer generated by this data.

[4]Credit to project partner Samuel J. Hall for these particular feature extractions.

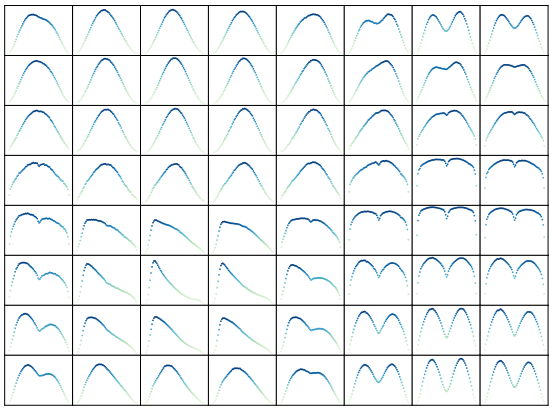

Figure 5: A restricted 8x8 projection of the underlying 40x40 Kohonen layer after 50 iterations. Each template plotted here represents the bottom left template of its 5x5 sub-grid. Note how the templates coincide with the clustering in Figure 4.

D. SOM Features

Using the most dominant period as calculated in section III.B., each lightcurve is then phase-folded, binned into 64 bins and normalised between 0 and 1. Bin values were calculated by taking the mean of points within its bin. These phase-folded lightcurves were then passed into the trained 2D SOM as in Figure 4. The SOM outputs 2D co-ordinates corresponding to the best matching template as shown in Figure 5 along with the squared residual distance to the best matching template. The outputted 2D coordinates are no use as data features to be inputted to the random forest. To combat this, each cluster in the trained SOM has been assigned a centre (e.g. RRab - $(11, 13)$); we then compute the distances to each of these and use them as a feature for the random forest.

E. The Random Forest Model

Here, we train and test the model on only the best 493 stars in our training set. A 75/25 train/test split and a random forest classifier with 500 estimators is used. The model correctly classifies 92.7% of the 123 test cases; this was the optimum score achieved in the given time frame. One should note that here, the classification of an input star is determined by the corresponding class with the largest probability output from the random forest which makes this a rather harsh metric to score the model. Perhaps a more useful summary of the model is shown in Figure 6. We observe promising behaviour, with most classifications falling on the diagonal of the matrix with high? probabilities. However, these outputted probabilities are subject to one final calibration step before they can be interpreted as true probabilities.

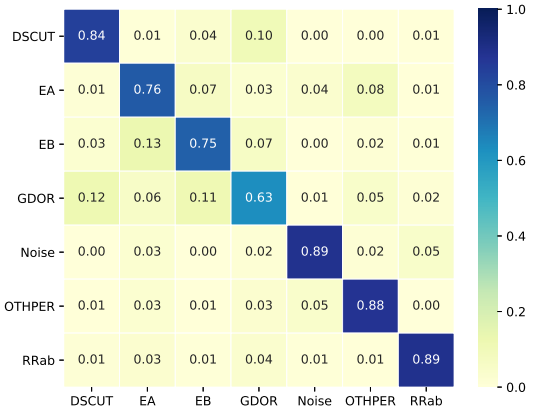

Figure 6: A confusion matrix for true vs. predicted classes of roughly 18 stars of each type. A given row shows the *average* probabilities outputted from the random forest. The row labels are the true classes.

IV. Plan for Term 2

The overall aim for Term 2 is to improve the accuracy and robustness of the model developed in Term 1. Once this has been achieved, we will use the final model to classify all stars in available K2 campaigns; producing a catalogue including probability estimates for each class. The anticipated steps and associated time frames to achieve this aim are as follows:

A. Larger Training Set (1 Week)

By choosing to use detrended data from K2SC, the available data for early campaigns is limited to campaigns 3 and 4 only[5]. Additionally, we have then restricted this data to the best classifications as determined in Armstrong et. al. 2016; comprising a test data set of roughly 500 stars. It is clear that this isn't large enough to gain an accurate insight to the model's underlying performance. Fortunately, there are other high-level science products (HLSP) available for the K2 campaigns which tackle the problem of removing the instrumental systematic noise. In particular, the EPIC Variability Extraction and Removal for Exoplanet Science Targets (EVEREST) data product has available lightcurves for campaigns 0-18. We hope to obtain a much larger test set spanning campaigns 0-4 using EVEREST data.

B. Other Classifiers (2 Weeks)

A simple extension to the RF classifier previously used is an *Extra Trees Classifier* which uses averaging to improve the predictive accuracy and control over-fitting. Other possible classifiers include AdaBoost, Gradient Boosting, and Naive Bayes; readily available in the `scikit-learn` package.

C. Deep Learning (2 Weeks)

Motivated by their ever-growing dominance and accuracy of image classification problems, we could perhaps improve our model by using a Neural Network instead of a SOM. Packages such as `tensorflow` will be explored and are readily available for such a problem.

D. Calibration, Analysis and Catalogue Production (1 Week)

The probabilities outputted from classifiers are fundamentally biased as explained in [26]. In particular, a Random Forest classifier has trouble with predictions of probabilities near 0 and 1 and so in order to present true probabilities in the final catalogue, calibration of the results is required. Classifications for stars in campaigns 5 and upwards will then be made[6].

[5] Campaigns 0, 1, 2 are yet to be processed by the K2SC team.

[6] Which campaigns will depend on the data product used as explained in Section IV.A.

References

[1] Richards, J. W. et al., The Astrophysical Journal **733** (2011) 10.

[2] Bloom, J. S. et al., Publications of the Astronomical Society of the Pacific **124** (2012) 1175.

[3] Neeley, J. R. et al., Monthly Notices of the Royal Astronomical Society **490** (2019) 4254.

[4] Riess, A. G. et al., The Astrophysical Journal **826** (2016) 56.

[5] Riess, A. G. et al., The Astrophysical Journal **730** (2011) 119.

[6] Bowman, D. M. et al., Astronomy and Astrophysics **621** (2019) A135.

[7] Lecoanet, D. et al., The Astrophysical Journal, Letters **886** (2019) L15.

[8] Ricker, G. R. et al., Journal of Astronomical Telescopes, Instruments, and Systems **1** (2015) 014003.

[9] Howell, S. B. et al., Publications of the Astronomical Society of the Pacific **126** (2014) 398.

[10] Gaia Collaboration et al., Astronomy and Astrophysics **595** (2016) A1.

[11] Gaia Collaboration, Astronomy and Astrophysics **616** (2018) A1.

[12] Armstrong, D. J. et al., Monthly Notices of the Royal Astronomical Society **456** (2016) 2260.

[13] Gautschy, A. and Saio, H., The Annual Review of Astronomy and Astrophysics **34** (1996) 551.

[14] Breger, M., Publications of the Astronomical Society of the Pacific **91** (1979) 5.

[15] McNamara, D. H., Madsen, J. B., Barnes, J., and Ericksen, B. F., The Publications of the Astronomical Society of the Pacific **112** (2000) 202.

[16] Guzik, J. A., Kaye, A. B., Bradley, P. A., Cox, A. N., and Neuforge, C., The Astrophysical Journal, Letters **542** (2000) L57.

[17] Smith, H. A., *RR Lyrae Stars*, Number **27** in Cambridge Astrophysics, Cambridge University Press, 2004.

[18] Catelan, M., Pritzl, B. J., and Smith, H. A., The Astrophysical Journal, Supplement **154** (2004) 633.

[19] Borucki, W. J. et al., Science **327** (2010) 977.

[20] Luger, R. et al., The Astrophysical Journal **152** (2016) 100.

[21] Aigrain, S. et al., Monthly Notices of the Royal Astronomical Society **459** (2016) 2408.

[22] Lomb, N. R., Astrophysics and Space Science **39** (1976) 447.

[23] Breiman, L., Machine Learning **45** (2001) 5.

[24] Richards, J. W. et al., The Astrophysical Journal **733** (2011) 10.

[25] Bailer-Jones, C. A. L. et al., The Astronomical Journal **156** (2018) 58.

[26] Niculescu-Mizil, A. and Caruana, R., ICML **01** (2005) 625.