MNRAS **000**, 1–8 (2018) Preprint 23 July 2018 Compiled using MNRAS LATEX style file v3.0

TESS Asteroseismic Predictions for Red Giants using Machine Learning

M. Schofield[1,2]★, G. R. Davies[1,2], W. J. Chaplin[1,2], A. Miglio[1,2]
[1]*Department of Physics and Astronomy, the University of Birmingham, Birmingham B15 2TT, UK*
[2]*Stellar Astrophysics Centre (SAC), Department of Physics and Astronomy, Aarhus University, Ny Munkegade 120, DK-8000 Aarhus C, Denmark*

Last updated 2017 May 31; in original form 2017 May 31

ABSTRACT

For the first time. a classifier was used as a way to select solar-like asteroseismic targets before a future mission. The classifier managed to identify the detectable solar-like oscillations inside *Kepler* Red Giants with a 0.98% precision. To do this, it used only the global parameters `[log(`g`), `π`, `$T_{\rm eff}$` , [M/H], `$I_{\rm mag}$` ]`.

The same classifier was also used on the *Kepler* stars after they had been degraded into TESS-like observations. The classifier scored 0.90% and 0.81% when made to look like 1-year and 27-days of TESS-like data, respectively. This classifier could be used to select asteroseismic targets for TESS in the Continuous Viewing Zone (Ricker et al. 2014), and as a way of investigating any target selection bias in previous missions such as K2 and CoRoT.

1 INTRODUCTION

- We use a classifier to make detectability predictions for TESS (Ricker et al. 2014)
- Mention the application to PLATO (Rauer et al. 2014).
- Supervised classifiers have been used to determine the evolutionary state of variable stars: Debosscher et al. (2007), Sarro et al. (2009), Nun et al. (2014), Elorrieta et al. (2016). Similarly, unsupervised classifiers (Valenzuela & Pichara 2018) and regression (Ness et al. 2015) have also been used to determine evolutionary stages.
- A neural network has also been used to identify the evolutionary stage of Red Giant stars (Hon et al. (2017), Hon et al. (2018)).
- Machine learning has also been used to calculate stellar parameters (Bellinger et al. 2016), including oscillation frequencies (Davies et al. 2016).
- One reason to use Machine Learning for target selection is that it can reverse-engineer target selection bias, for example in K2 (Lund et al. 2015) and CoRoT (Baglin et al. 2006). This can provide insights into the formation history of the galaxy: Thomas et al. (2017).
- Summarise the Chapters.

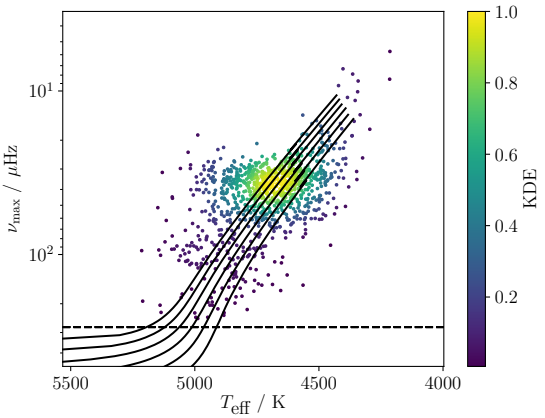

Figure 1. The 1000 *Kepler* Red Giants from Davies et al. (in prep). The colourbar shows the relative density of points with Kernel Density Estimation. The evolutionary tracks range from 0.9-1.5$M_\odot$.

2 THE DATASET

The data to perform Machine Learning on are the 1000 *Kepler* Red Giants from Davies et al. (in prep). These are shown in Figure 1. The radial and quadrapole modes of these stars have been fitted by Davies et al. (in prep). The frequency, height, width and background of each mode in the 1000 stars had been individually fitted. The spectroscopic parameters of these stars ($T_{\rm eff}$, log(g) , [M/H]) were available from APOKASC Pinsonneault et al. (2014). The apparent magnitudes are from the *Tycho*-2 catalogue Hog et al. (2000). Lastly, the parallaxes are from *GaiaDR*2 Lindegren et al. (2018).

Machine Learning performs best when a large dataset is available to train the algorithm on. In order to increase the size of the dataset above 1000, the magnitude of each *Kepler* star was perturbed. Each star had its apparent magnitude perturbed 100 times.

The instrumental (shot) noise models of *Kepler* and TESS were used as PDFs to draw apparent magnitudes from. These were used because they provide realistic distributions of the number of stars at different magnitudes that the satellites observed/will observe. Many more fainter stars are observed than brighter stars because the volume of space that contains stars increases as the distance of observation increases.

The shot noise model of *Kepler* depends on the *Kepler* magnitude of the star, K_p. This noise model is from Gilliland et al.

(2010). The RMS noise, σ, is given by

$$\sigma = \frac{10^6}{c} \times \sqrt{c + 9.5 \times 10^5 \left(\frac{14}{K_p}\right)^5} \text{ ppm}, \quad (1)$$

where c is the number of detected electrons per cadence. It is given by

$$c = 1.28 \times 10^{0.4(12-K_p)+7}. \quad (2)$$

The equivalent TESS shot noise model was taken from Sullivan et al. (2015). The RMS noise was calculated using photon counting noise, the noise from background stars, the readout noise and the systematic noise. These four noise sources were then summed in quadrature to give the total TESS noise.

The *Kepler* and TESS noise models were used as the PDFs to draw stellar magnitudes from. 100 magnitudes were drawn for every *Kepler* Red Giant. After removing gaps in the data, this left 60,000 samples. The lightcurves of this much larger *Kepler* dataset were then degraded to look like TESS observations.

3 IMPUTING IMAG VALUES

741 of the 1000 *Kepler* stars have measured $I_{\rm mag}$ values from *Tycho* − 2 (Hog et al. 2000). $I_{\rm mag}$ is needed for every star to calculate the TESS shot noise level. In order to keep the remaining stars in the dataset, the $I_{\rm mag}$ values were imputed.

$I_{\rm mag}$ values for the missing stars were imputed using random forest regression. Like a random forest classifier, random forest regression is another example of supervised learning: in both cases there is a known label to predict (here the label is $I_{\rm mag}$). Supervised learning involves separating the data into training and testing datasets. In classification, the data is grouped by similarity. In regression, the difference between the regression model and 'true' values is evaluated, and iteratively reduced. This difference is evaluated using the Sum of Squared Errors (SSE);

$$\text{SSE} = \sum_{i=1}^{N} (x_i - \bar{x})^2. \quad (3)$$

Random forest regression was used in this way to get the missing $I_{\rm mag}$ values. This could have been done using only the 741 available $I_{\rm mag}$ values of the 1000 stars. However, in order to improve the predictions made from the regression, a selection of the brighter stars in the *Kepler* Input Catalogue (KIC; Brown et al. (2011)) were used alongside the 741 fitted Red Giants. In total, 2366 stars were used to train and test the regression.

The KIC provides a variety of magnitudes, but not the 2MASS[1] $I_{\rm mag}$. $I_{\rm mag}$ was calculated for the KIC stars using the 2MASS magnitudes $J_{\rm mag}, H_{\rm mag}, K_{\rm mag}$, and the SDSS (Kollmeier et al. 2017) magnitude $i_{\rm mag}$. Not all of these magnitudes were available for the 259 fitted *Kepler* stars from Davies et al. (in prep).

Firstly, an equation from Bilir et al. (2008) was used to calculate $(R - I)$;

$$(R - I) = c_1(J - H) + c_2(H - K) + c_3. \quad (4)$$

The coefficients c_1, c_2 and c_3 vary with metallicity and are given in the paper. Secondly, $(i - I)$ was calculated from $(R - I)$ using an equation from Jordi et al. (2006);

$$(i - I) = 0.247(R - I) + 0.329. \quad (5)$$

[1] https://www.ipac.caltech.edu/2mass/

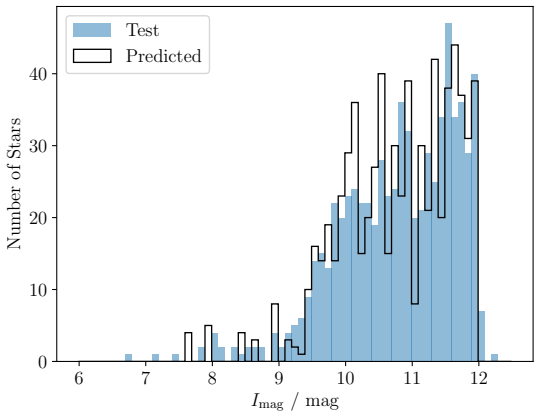

Figure 2. The $I_{\rm mag}$ distribution of the KIC stars used to test the random forest regression. The 'true' values used to test the algorithm are shown in blue. The black histogram shows the distribution of values that the random forest regression predicted for that dataset.

It is then straightforward to calculate $I_{\rm mag}$ for every KIC star,

$$I_{\rm mag} = -(i - I) + i_{\rm mag}. \quad (6)$$

Once $I_{\rm mag}$ values were calculated for every KIC star in the dataset, they were used in random forest regression along with [K_p, [M/H], $T_{\rm eff}$] to predict $I_{\rm mag}$ value for the 259 *Kepler* stars.

70% of the dataset was used to train the algorithm; 30% was used to test it's accuracy. The results of the random forest regression are shown in Figures 2 and 3. Figure 2 shows that the distribution of predicted value matches the 'true' $I_{\rm mag}$ values closely. Figure 3 shows that there is no offset between the predicted and 'true' values; the mean difference between the two is -0.01 mag. Furthermore, the standard deviation of the difference is only 0.27 mag, and the regression achieves an accuracy of 0.90. To summarise this, random forest regression predicts the $I_{\rm mag}$ values in the test dataset well.

After testing the regression, the algorithm was used to make predictions on the *Kepler* stars without known $I_{\rm mag}$ values. The distribution of predicted values is shown in Figure 4, along with the entire KIC sample used to train the regression in blue for comparison.

4 MODIFYING KEPLER DATA TO LOOK LIKE TESS

Before a classifier could be used on the Red Giant sample, the time-series data from *Kepler* needed to be modified for TESS. These adjustments were made in the time domain before the signal was converted to the frequency domain.

Several different adjustments needed to be made to the *Kepler* data. One difference between the missions is the length of observation. The *Kepler* mission observed stars for up to 4 years. TESS' nominal 2 year mission will observe stars for between 27 days to 1 year, according to the ecliptic latitude of the stars (Ricker et al. 2014).

As well as reducing the dataset length, the bandpass of observation needed to be adjusted. TESS will observe in a much redder bandpass than that of *Kepler* . This has the effect of reducing the amplitude of stellar signals (i.e the signals due to stellar granulation and solar-like oscillations) (Ballot et al. 2011). Campante et al.

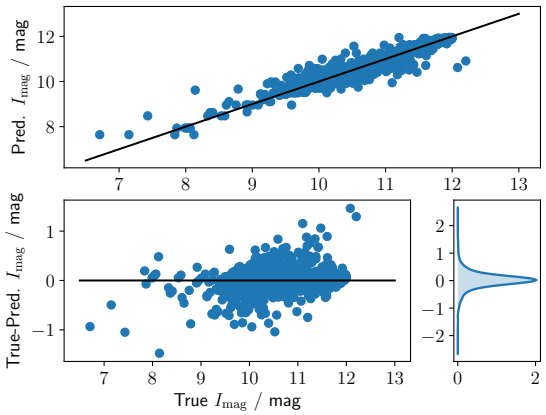

Figure 3. The true I_{mag} values of the KIC stars use to test the algorithm, compared to their predicted values. The mean difference between the two sets of values is -0.01 mag, with a standard deviation of just 0.27 mag.

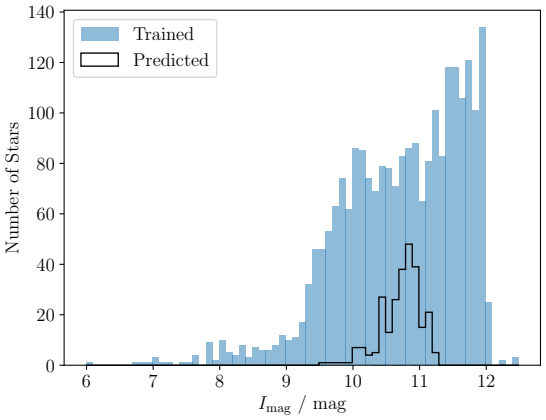

Figure 4. The I_{mag} distribution of the KIC stars used to train the random forest regression is shown in blue. The predicted values for the stars without I-band magnitudes is shown in black.

(2016) found this bandpass correction in the amplitude spectrum to be 0.85 for TESS.

Thirdly, the noise level for a given star in *Kepler* is lower than the noise level in TESS. Noise from TESS was calculated for each star and added to the timeseries (see Section 2).

These three adjustments - the length of observation, the bandpass, and the noise level (see Section 2) - were performed in the time domain. From the original *Kepler* timeseries, these adjustments were made:

(i) Apply a 4-σ clip to the dataset to remove spurious points.
(ii) Cut the 4 years of data down to the reduced dataset length (27 days to 1 year). Use the 27 days to 1 year of data with the fewest gaps.
(iii) Adjust the bandpass by multiplying the flux by 0.85.
(iv) Add TESS instrumental noise to the timeseries. To do this, calculate the TESS `RMS` noise level (Section 2). For each flux value in the timeseries, draw a random number from the normal distribution. Multiply the `RMS` noise by each flux value. Add this to the original flux values in the timeseries.

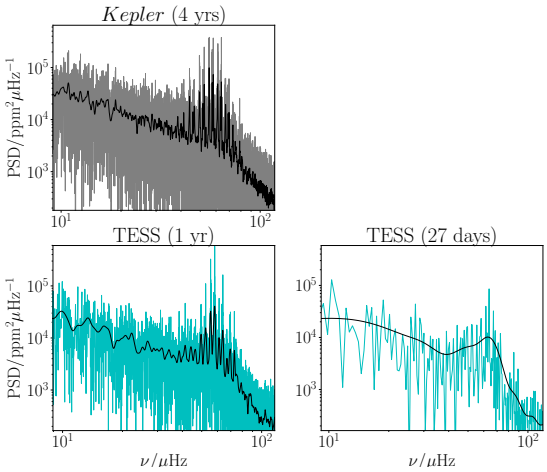

Figure 5. The Power spectra of KIC 9535399 is plotted three times with moving medians in black. The original Power Spectra is plotted in grey. The power spectra after making the data look like TESS are plotted in blue. The subplot on the bottom left shows 1 year of TESS observation (the maximum). The subplot on the bottom right column shows 27-days (the minimum).

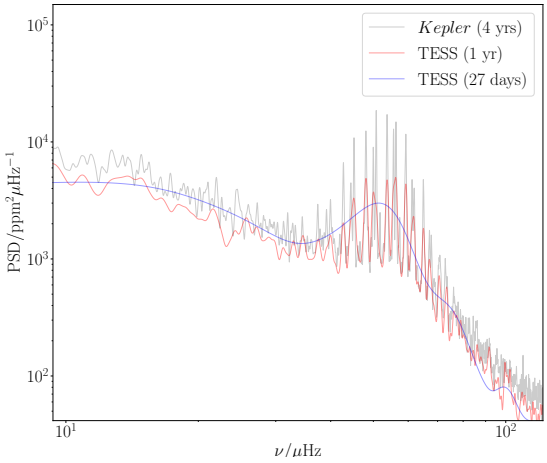

Figure 6. The Power Spectra of KIC 6768319. The original power spectra is in grey. The data were transformed into 1 year and 27 days of TESS observation. The moving median of these transformations are overplotted.

(v) Use the Lomb-Scargle periodogram to convert the signal from the time to frequency domain.

Examples of the TESS-like power spectra are shown in Figures 5 and 6. Power spectra were generated for the original 4-year *Kepler* sample, 1 year of TESS observations and 27 days of TESS observations. Once the timeseries had been made to look like TESS observations, a detection test was then run on the radial modes of every star in these 3 datasets.

5 DETECTION TEST

Section 4 described the method to transform the *Kepler* lightcurves into TESS-like power spectra. A detection test was then run on the stars to determine which modes were detectable by TESS, and which were not.

First, a moving median was used to estimate the underlying

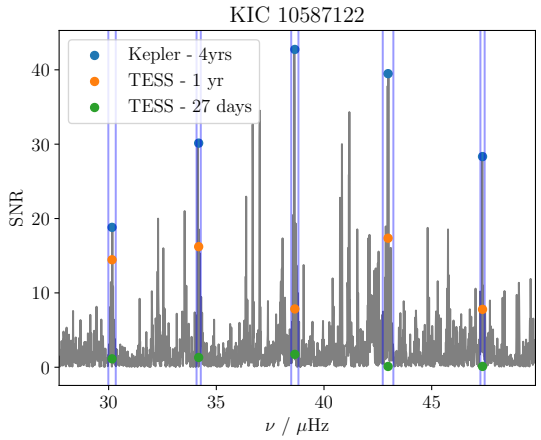

Figure 7. The SNR spectrum of KIC 10587122 after background subtraction. The SNR values of the radial modes in the star were extracted from this spectrum. The highest SNR value within the linewidth of each mode is taken to be the SNR value of that mode. The mode linewidths are shown as blue lines. The values of every mode in the original Kepler spectrum are plotted as blue points. The overplotted orange points are the SNR values after degrading the signal to 1 year of TESS observation. Similarly, the green points are the SNR values of 27 days of TESS observations. The white noise level and reduced observation time severely reduce the SNR of TESS observations compared to *Kepler* .

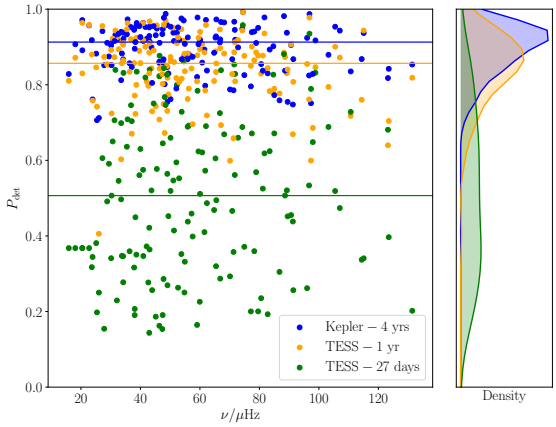

Figure 8. A plot showing the result of the detection test, after running on every mode in 20 stars. The results of the original power spectra are plotted in blue. The results of 1 year of TESS observation are in orange. 27 days of TESS observation is in green. At this short an observation, detecting individual modes will be extremely difficult.

background spectrum. The solar-like mode envelope width was used as the frequency range of this moving median. This envelope width was calculated as

$$\Gamma_{\rm env} = 0.66\,\nu_{\rm max}^{0.88}, \tag{7}$$

from Mosser et al. (2012). The moving median provided an estimate of the background B in the power spectrum (ppm$^2\mu$Hz^{-1}). This background was divided out of the power P (ppm$^2\mu$Hz^{-1}) in the power spectrum to the get Signal-to-Noise ratio of the spectrum,

$$\mathrm{SNR} = P/B. \tag{8}$$

Once the SNR spectrum for the star was recovered, the SNR values at the mode frequencies were extracted. To ensure the correct SNR values of every mode were used, a window was fitted around each mode frequency from Davies et al. (in prep). The size of the window was given as the linewidth of the mode. The highest value in the window was taken as the SNR of the mode. An example of this for KIC 10587122 is shown in Figure 7.

Once all mode SNR values for the star were calculated, a detection test from Chaplin et al. (2011) was run on each mode. In Chaplin et al. (2011), a detection test was used across the entire oscillation envelope. Here, the same detection test was instead applied to individual modes. This method is described below.

The probability P that the SNR value of a solar-like mode of oscillation lies above some threshold $\mathrm{SNR}_{\rm thresh}$ is

$$P(\mathrm{SNR} \geqslant \mathrm{SNR}_{\rm thresh}) = p. \tag{9}$$

A false-alarm probability p of 5% was set; there is a 95% chance that the signal is not due to noise. Equation 9 is solved for $\mathrm{SNR}_{\rm thresh}$ by substituting P with

$$P = \int_x^\infty \frac{e^{-x}}{\Gamma(N)} x^{N-1} dx. \tag{10}$$

N is the number of frequency bins that the mode occupies. The linewidth of each mode was used as the value of N in equation 10.

$\Gamma(N)$ is the Gamma function. The lower bound of Equation 10 is set to $x = 1 + \mathrm{SNR}_{\rm thresh}$. The noise in the N bins is assumed to follow χ^2 $2n_{\rm bins}$ d.o.f statistics.

Once $\mathrm{SNR}_{\rm thresh}$ is found, Equation 10 is solved again. This time it is solved for P by setting $x = (1 + \mathrm{SNR}_{\rm thresh})/(1 + \mathrm{SNR})$. Here, SNR is the observed Signal-to-noise Ratio calculated from Equation 8. This calculates the probability that the frequency spike was due to stochastic excitation in the convective envelope of the star.

This detection test was applied to every mode of every star in the sample. Figure 8 shows the mode detection probabilities from this test for the original *Kepler* dataset, for 1-year of TESS observation and for 27 days of TESS observation. After the $P_{\rm det}$ values for the 3 datasets were calculated, a classifier was used to see if these results could be predicted rather than calculated. This is described in Section 6.

6 CLASSIFICATION

Section 4 describes how the lightcurves of every star were treated before a detection test was run on the oscillations in Section 5. After this, a random forest classifier was used to predict the detection probability of the modes in the Red Giants.

Classification algorithms work by assessing similarity[2]. In Classification, the training set is separated into groups based on the similarity of the data. The more information that was gained by splitting the data, the better. The data continues to be split until a prediction can be made (in this case about the detection probabilities of 3 radial modes). In decision tree classification, this is done once.

Here, a random forest classifier was used. This is made up of many independent Decision Trees which have been weighted to predict the detection probabilities of the 3 radial modes. Section 6.1 will explain how the data was prepared before the random forest classifier was used.

[2] http://www.simafore.com

KIC	Iteration	log(g)	π	$T_{\rm eff}$	[M/H]	$I_{\rm mag}$
9205705	1	2.758	0.688	4685	-0.39	9.89
9205705	2	2.758	0.688	4685	-0.39	9.19
9205705	3	2.758	0.688	4685	-0.39	9.79
9205705	4	2.758	0.688	4685	-0.39	11.30
...						
9205705	100	2.758	0.688	4685	-0.39	7.81
2554924	1	2.799	0.969	4594	0.27	8.46
2554924	2	2.799	0.969	4594	0.27	9.26
...						

Table 1. An example of the X-dataset for 1 year of TESS-like observations. Every star has it's magnitude perturbed 100 times. See Table 2 for the equivalent Y-dataset.

KIC	Iteration	$P_{\rm det}$ (1)	$P_{\rm det}$ (2)	$P_{\rm det}$ (3)
9205705	1	1	2	2
9205705	2	1	2	2
9205705	3	1	2	2
9205705	4	1	2	1
...				
9205705	100	1	2	2
2554924	1	2	2	2
2554924	2	2	2	2
...				

Table 2. An example of the Y-dataset for 1 year of TESS-like observations. Every star has it's magnitude perturbed 100 times. White noise is then added to the timeseries and mode detection probabilities are calculated for 3 radial modes centred around $\nu_{\rm max}$. Lastly, these probabilities are put into discrete classes [0, 1 or 2]. The radial mode closest to $\nu_{\rm max}$is $P_{\rm det}$ (2). See Table 1 for the equivalent X-dataset.

6.1 Preparing the data

After removing stars without APOKASC information, 811 stars were left from the original 1000-star dataset. In Section 2, each star had it's $I_{\rm mag}$ value perturbed 100 times. After perturbing the apparent magnitudes, 81,100 stars were available to perform classification upon. This was done 3 times: when the stars were treated as *Kepler* targets, when they were treated as TESS targets in the Continuous Viewing Zone (1 year of observation), and when the 81,100 stars were treated as TESS targets observed for 27 days.

Each calculated detection probability from Section 5 was then put into a discrete bin (or *class*) depending on how likely the mode is to be detected. These discrete classes are given in equation 11.

$$P_{\rm det} = \begin{cases} 2 & \text{if } 1.0 \geqslant P_{\rm det} > 0.9 \ , \\ 1 & \text{if } 0.9 \geqslant P_{\rm det} > 0.5 \ , \\ 0 & \text{if } 0.5 \geqslant P_{\rm det} > 0.0 \ . \end{cases} \tag{11}$$

Using equation 11, every mode was assigned a discrete class [0, 1 or 2], depending on how high the probability of detection was for that mode. The same three radial modes (*3 features*) were used for every star: the mode closest to the centre of the power-excess due to solar-like oscillations $\nu_{\rm max;n}$, the radial mode one overtone below that $\nu_{\rm n-1}$, and one overtone above that, $\nu_{\rm n+1}$ [$P_{\rm det}$ (1), $P_{\rm det}$ (2), $P_{\rm det}$ (3)]. It was important to use the same modes for every star so that the algorithm could be trained on the patterns between the variables.

A classifier is an algorithm that can learn a relationship between variables. The classifier will map from some initial information about the star (the X data), to some unknown information (the Y data). In this work, the X data features were magnitude (K_p or $I_{\rm mag}$), log(g), π, $T_{\rm eff}$ and [M/H]. log(g) , $T_{\rm eff}$ and [M/H] values are from Pinsonneault et al. (2014), π is from *Gaia* DR2 Lindegren et al. (2018) and $I_{\rm mag}$ is from Hog et al. (2000), with some values imputed using regression (Section 3).

The Y data features were the $P_{\rm det}$ values of 3 radial modes centred around $\nu_{\rm max}$. An example of the final dataset for 1 year of TESS-like observations are shown in Tables 1 and 2.

6.2 Target selection using a classifier

The 81,100 datasets were separated into a training dataset, and a testing set. 70% of the datasets were used to train the classifier (56,770 stars); 30% of the stars were used to test the algorithm (24,330 stars). To train the classifier, the X and Y data in the training set was given to the algorithm ($\rm X_{train}$ and $\rm Y_{train}$). Once the classifier had been trained, the X data from the testing set was given to it ($\rm X_{test}$). The classifier then predicted a set of Y data for the testing set ($\rm Y_{pred}$). This was compared to the actual Y data for the testing set ($\rm Y_{test}$). The more similar $\rm Y_{pred}$ is to $\rm Y_{test}$, the better the classifier replicated the data.

Two metrics were used to measure the performance of the algorithm. The first was the precision of the classifier. To calculate the precision, the difference between the 'true' values and the values predicted by the classifier were calculated. This was done for each feature [$P_{\rm det}$ (1), $P_{\rm det}$ (2), $P_{\rm det}$ (3)] separately. The mean of these differences was calculated, and weighted by the number of 'true' values in each feature.

This precision $P_{\rm res}$ is given as

$$P_{\rm res} = t_{\rm p}/(t_{\rm p} + f_{\rm p}), \tag{12}$$

where $t_{\rm p}$ are true-positives and $f_{\rm p}$ are false-positives. The classifier's precision is its ability to not label a negative sample as positive[3].

The second was the Hamming loss[3] (Wegner 1960) of the algorithm. This was used to give another measure of similarity between the predicted $P_{\rm det}$ values $\rm Y_{pred}$, and the testing $P_{\rm det}$ values $\rm Y_{test}$;

$$H_{\rm loss}(Y_{\rm test}, Y_{\rm pred}) = \frac{1}{n_{\rm classes}} \sum_{j=0}^{n_{\rm classes}-1} 1(Y_{\rm pred} \neq Y_{\rm test}) \ . \tag{13}$$

A Hamming loss score of 0.0 means that $\rm Y_{pred}$ is identical to $\rm Y_{test}$. A score of 1.0 means that there are no similar values between $\rm Y_{pred}$ and $\rm Y_{test}$. Precision and Hamming loss are similar (but not identical) measures of the success of a classifier. By using both metrics, the usefulness of the classifier can be better assessed. The precision and Hamming loss of the classifier on the *Kepler* and TESS datasets are shown in Table 3.

We also tested the impact of increasing the number of classes in equation 11, and the range of $P_{\rm det}$ values for each class. The number of classes was varied from 2 (i.e the mode was detected (1) or it was not (0)) to 6. The width of each bin was also varied to ensure that bins were not underpopulated. It was found that the 3 classes and $P_{\rm det}$ ranges given in equation 11 gave the best predictions for the 3 datasets of 81,100 stars (4 years of *Kepler* observation, 1 year of TESS observation and 27 days of TESS observation).

The results for the original *Kepler* dataset and for 1-year of

[3] http://scikit-learn.org

Satellite	T_{obs}	Precision	Hamming loss
Kepler	4 years	0.98	0.02
TESS	1 year	0.90	0.09
TESS	27 days	0.81	0.19

Table 3. Results of the classifier on the original *Kepler* dataset, and the 1-year and 27-day TESS datasets. The 'Precision' column gives the average weighted precision of the classifier across the 3 classes [0, 1, 2] and 3 features [P_{det} (1), P_{det} (2), P_{det} (3)].

TESS data are very good; the classifier was able to replicate the mode detection predictions of the stars in these cases. This means that a classifier can be used as a tool for target selection for future missions. Once the classifier predicted P_{det} values, the stars could be ranked from those with many detected modes, to those with the fewest. In this way, the classifier could be used as the target selection function of solar-like oscillators for TESS.

For the 27-day TESS dataset, the classifier was able to correctly predict the P_{det} values of 81% of the modes. This is likely to be because with these TESS targets, the dataset is too susceptible to the realization noise of the observation for individual modes to be reliably detected. A mode may be clearly detectable in one 27-day observation, and not another.

6.3 Feature Importance

Here, a classifier was used to predict the detection probabilities of individual modes in *Kepler* Red Giant stars. As well as being much faster than a conventional mode detection test, an added bonus of the classifier method is that it returns the 'feature importance' of each label in the X-data.

As Table 1 shows, the X-data features that are given to the classifier are `[log(g), π, T_eff , [M/H], I_mag ]`. Some of these features are more important than others when predicting the detection probability of solar-like modes `[P_det (1), P_det (2), P_det (3)]`. The relative importance of these columns is known as the feature importance. This feature importance sums to 1.

The feature importances of the X-data columns are given in Figure 9. The feature with the highest influence on mode detectability is apparent stellar magnitude I_{mag} . This is not surprising: a higher value of I_{mag} will result in more χ^2 2 Degrees-of-Freedom white noise in the observation. This will make oscillations less likely to be detected above the white noise background.

After the apparent stellar magnitude, surface gravity log(g) also has a heavy influence on P_{det} value. This is because log(g) is a proxy for the central frequency of the solar-like mode envelope ν_{max} ($\nu_{max} \propto \log(g)$). As a star evolves from the Main-Sequence and up the Red Giant Branch, the radius of the star increases. This decreases the log(g) and ν_{max} values of the star. Kjeldsen & Bedding (1995) showed that oscillation amplitude is proportional to bolometric luminosity divided by mass ($A_{osc} \propto L/M$). As a star evolves 'up' the Hertzsprung-Russell diagram, its bolometric luminosity increases while it's mass stays roughly the same. This leads to larger oscillation amplitudes which are more likely to be detected. Mathur et al. (2011) also gives a good explanation of how the granulation properties (and hence the oscillation profile) change as a star evolves.

Comparatively, effective temperature T_{eff} is less useful than log(g) when predicting detection probability. It is less closely tied to the oscillation properties of the star than log(g) , although it is a property used to estimate ν_{max} with the scaling relations (Chaplin et al. 2011).

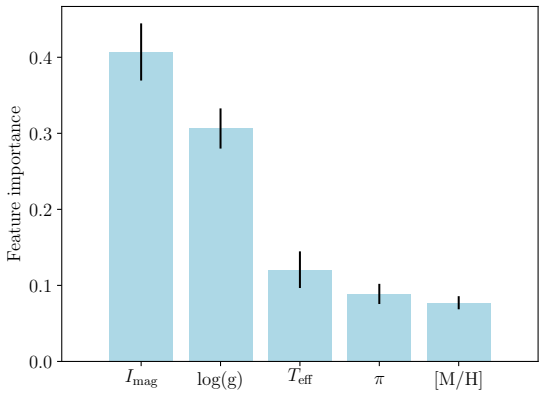

Figure 9. The feature importance of the 5 X-data columns.

Lastly, parallax π and metallicity [M/H] are the least informative features when predicting the detection probabilities of these stars. Parallax is connected to apparent magnitude and T_{eff} , while [M/H] is connected to log(g). The feature importances of T_{eff} , π and [M/H] highlight the additional information that these values bring, as well as indirectly through I_{mag} and log(g) .

6.4 Comparing results between different evolutionary states

So far in this paper, the the evolutionary state of the Red Giant population from Davies et al. (in prep) has been ignored when predicting mode detection probability P_{det}. In reality, the 1000 *Kepler* Red Giants used in this work are a mixture of Red Giant Branch (RGB), Red Clump (RC) and Secondary Clump (2CL) stars. This Section investigates whether there is any difference in the predictions if the stars are first separated into RGB, RC and 2CL groups. The evolutionary states of the stars in the sample were taken from Elsworth et al. (2017).

If a Red Giant Branch is massive enough, it will undergo the Helium flash and become a Red Clump star. Red Clump stars are Helium-core burning stars, and all have very similar core masses to each other. These stars have very different $g-$mode period spacings to RGB stars, but are otherwise difficult to differentiate (Chaplin & Miglio (2013), Bedding (2011), Beck et al. (2011)). If a star is more massive than $\gtrapprox 1.8 M_{\odot}$, it will instead become a Secondary Clump star. This means that it will undergo Helium-core burning without the Helium flash.

The Red Giant dataset was separated into 3 subsets; the RGB, RC and 2CL stars. Each subset was treated in the same way as in Table 3. Table 4 gives the full list of results when the RGB, RC and 2CL stars are separated.

Table 4 shows that there is a negligible difference in predictions between stars undergoing Helium-core burning (RC and 2CL stars) and those that are not (RGB stars). This leads to the conclusion that a classifier can predict the mode detectability of Red Giants at different evolutionary states equally well. It is therefore not necessary to separate out these stars into different evolutionary states before predicting P_{det} .

Evolution	Satellite	T_{obs}	Precision	Hamming loss
RGB	*Kepler*	4 years	0.98	0.03
RGB	TESS	1 year	0.83	0.16
RGB	TESS	27 days	0.75	0.25
RC	*Kepler*	4 years	0.96	0.04
RC	TESS	1 year	0.90	0.09
RC	TESS	27 days	0.77	0.24
2CL	*Kepler*	4 years	0.97	0.03
2CL	TESS	1 year	0.83	0.14
2CL	TESS	27 days	0.69	0.30

Table 4. Results of the classifier when RGB, RC and 2CL stars are separated. Results are shown when the data are treated like *Kepler* stars, and when they are degraded to look like 1-year and 27-day TESS observation. The 'Precision' column gives the average weighted precision of the classifier across the 3 classes [0, 1, 2] and 3 features [P_{det} (1), P_{det} (2), P_{det} (3)].

7 CONCLUSION

The solar-like oscillations of 1000 *Kepler* Red Giant stars from Davies et al. (in prep) were used to determine whether asteroseismic target selection could be done using a classifier. Rather than using Machine Learning to determine the evolutionary state of these stars, a classifier was instead used to predict which individual solar-like oscillations inside the Red Giants could be detected with a future space mission. The mission in question here was TESS, although the same technique can be easily applied to other missions such as PLATO. This tool can also be used to understand target selection bias in previous missions, such as K2 and CoRoT.

Firstly, the number of samples was increased by perturbing stellar magnitudes 100 times for each star. These perturbed magnitudes were drawn from a PDF of the noise function (Section 2). After removing stars where global parameters or fitted modes were unavailable, this left 60,000 *Kepler* samples.

Once the number of samples was increased, the timeseries' were degraded to transform them into TESS-like observations (Section 4). The dataset length was reduced, white noise was added to the signal, and the bandpass of observation was reddened. A moving median was calculated for the power spectra of these TESS-like Red Giants to estimate the total background in the signal. This was divided out of the spectra, leaving a Signal-to-Noise ratio at every frequency bin (equation 8).

A detection test was then run on the SNR values at every mode frequency (Section 5). This gave a detection probability P_{det} between 0.0 and 1.0 for every mode. In order to prepare the detection probabilities before Classification, each continuous P_{det} value was assigned a discrete class ([0,1 or 2]; equation 11).

A classifier was then given the global photometric and spectroscopic properties of the Red Giant sample, along with mode detection probabilities for each star. The parameters `[log(g), π, T_eff , [M/H], I_mag ]` from APOKASC (Pinsonneault et al. 2014), TGAS (Gaia Collaboration et al. 2016) and *Tycho*-2 (Hog et al. 2000) were the 5 X-data features. The P_{det} values of 3 radial modes centred around ν_{max} were used as the Y-data features. The classifier used the global stellar parameters (the X-data) to make predictions about mode detectability (the Y-data). The stars with the largest number of detected modes could then be selected as the Red Giants for observation by TESS.

The classifier successfully made predictions about the original 4 years of Kepler data; the algorithm had a weighted precision 0.98 across the 3 P_{det} features. This confirms the proof of concept that classifiers can be used as a way to select solar-like asteroseismic targets before future missions. As well as this, this shows that it can be used to investigate any possible target selection bias, for example in K2 or CoRoT. Classification vastly reduces the computation time required to produce a target selection function, especially when large datasets are involved ($\geqslant 50,000$ stars).

Degrading the Red Giant data to make predictions for 1 year of TESS observations was also successful. The predicted mode detections scored a weighted precision of 0.90 across the 3 P_{det} features. This illustrates that Classification is a valid target selection method for TESS targets in the Continuous Viewing Zone (CVZ, (Ricker et al. 2014)).

Using the classifier on 27 days of TESS data returned detection predictions with a precision of 0.81. This is too low for the classifier to be used to select solar-like oscillators for 27 days of TESS observation. The precision is lower when stars are observed by TESS for 27 days because the white noise level is too high and the length of observation is too short to make robust predictions of individual solar-like oscillations. It may be that individual solar-like oscillations cannot be detected in 27 days of TESS data. If this is the case, then the classifier should not be expected to make robust detection predictions for these stars.

REFERENCES

Baglin A., et al., 2006. p. 3749, http://adsabs.harvard.edu/abs/2006cosp...36.3749B

Ballot J., Barban C., van't Veer-Menneret C., 2011, Astronomy and Astrophysics, 531, A124

Beck P. G., et al., 2011, Science, 332, 205

Bedding T. R., 2011, arXiv:1107.1723 [astro-ph]

Bellinger E. P., Angelou G. C., Hekker S., Basu S., Ball W. H., Guggenberger E., 2016, The Astrophysical Journal, 830, 31

Bilir S., Ak S., Karaali S., Cabrera-Lavers A., Chonis T. S., Gaskell C. M., 2008, Monthly Notices of the Royal Astronomical Society, 384, 1178

Brown T. M., Latham D. W., Everett M. E., Esquerdo G. A., 2011, The Astronomical Journal, 142, 112

Campante T. L., et al., 2016, preprint, 1608, arXiv:1608.01138

Chaplin W. J., Miglio A., 2013, Annual Review of Astronomy and Astrophysics, 51, 353

Chaplin W. J., et al., 2011, The Astrophysical Journal, 732, 54

Davies G. R., et al., 2016, Monthly Notices of the Royal Astronomical Society, 456, 2183

Debosscher J., Sarro L. M., Aerts C., Cuypers J., Vandenbussche B., Garrido R., Solano E., 2007, Astronomy and Astrophysics, 475, 1159

Elorrieta F., et al., 2016, Astronomy and Astrophysics, 595, A82

Elsworth Y., Hekker S., Basu S., R. Davies G., 2017, Mon Not R Astron Soc, 466, 3344

Gaia Collaboration et al., 2016, Astronomy and Astrophysics, 595, A2

Gilliland R. L., et al., 2010, The Astrophysical Journal Letters, 713, L160

Hog E., et al., 2000, Astronomy and Astrophysics, 355, L27

Hon M., Stello D., Yu J., 2017, Monthly Notices of the Royal Astronomical Society, 469, 4578

Hon M., Stello D., Yu J., 2018, Monthly Notices of the Royal Astronomical Society

Jordi K., Grebel E. K., Ammon K., 2006, Astronomy and Astrophysics, 460, 339

Kjeldsen H., Bedding T. R., 1995, Astronomy and Astrophysics, 293, 87

Kollmeier J. A., et al., 2017, preprint, 1711, arXiv:1711.03234

Lindegren L., et al., 2018, preprint, 1804, arXiv:1804.09366

Lund M. N., Handberg R., Davies G. R., Chaplin W. J., Jones C. D., 2015, The Astrophysical Journal, 806, 30

Mathur S., et al., 2011, The Astrophysical Journal, 741, 119

Mosser B., et al., 2012, A&A, 537, A30

Ness M., Hogg D. W., Rix H.-W., Ho A. Y. Q., Zasowski G., 2015, The Astrophysical Journal, 808, 16
Nun I., Pichara K., Protopapas P., Kim D.-W., 2014, The Astrophysical Journal, 793, 23
Pinsonneault M. H., et al., 2014, The Astrophysical Journal Supplement Series, 215, 19
Rauer H., et al., 2014, Experimental Astronomy, 38, 249
Ricker G. R., et al., 2014, Journal of Astronomical Telescopes, Instruments, and Systems, 1, 014003
Sarro L. M., Debosscher J., LÃşpez M., Aerts C., 2009, Astronomy and Astrophysics, 494, 739
Sullivan P. W., et al., 2015, The Astrophysical Journal, 809, 77
Thomas A., Stevenson E., Gittins F. W. R., Miglio A., Davies G., Girardi L., Campante T. L., Schofield M., 2017. eprint: arXiv:1610.08862, p. 05006, doi:10.1051/epjconf/201716005006, http://adsabs.harvard.edu/abs/2017EPJWC.16005006T
Valenzuela L., Pichara K., 2018, Monthly Notices of the Royal Astronomical Society, 474, 3259
Wegner P., 1960, Commun. ACM, 3, 322

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