MNRAS **000**, 1–9 (2018) Preprint 1 August 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

In this work, a classifier was used to make predictions about the detectability of solar-like modes of oscillation. Specifically, predictions of mode detectability for *Kepler* Red Giant Branch stars were made. After this, these *Kepler* Red Giants were degraded to produce a TESS-like dataset. Mode detectability predictions were then made using the same classifier on these TESS-like observations.

In the past, 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. As well as with classifiers, a neural network has been used to identify the evolutionary stage of Red Giant stars (Hon et al. (2017), Hon et al. (2018)).

As well as identifying the evolutionary stage of stars, Machine Learning has been used to calculate stellar parameters (Bellinger et al. 2016) and oscillation frequencies (Davies et al. 2016).

For the first time, this work makes predictions about the detectability of solar-like modes inside *Keper* and TESS targets using a classifier. This same technique can be applied to any future space mission, such as PLATO (Rauer et al. 2014). As well as being able to apply the same technique to any future mission, Machine Learning has another advantage as a target selection tool. Machine Learning is also useful for target selection because 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).

Firstly, Section 2 describes the *Kepler* Red Giant dataset from Davies et al. (in prep). These 1000 stars have measured values for the global properties [log(g), π, $T_{\rm eff}$, [M/H], $I_{\rm mag}$], as well as fitted radial mode frequencies, heights and widths. Most (but not all) of these Red Giant stars have been observed in the I-band. Section 3 describes how regression was used to infer the $I_{\rm mag}$ values of the missing stars, using the rest of the Red Giant dataset.

After the *Kepler* dataset was prepared, the stars were modified to mimic observations by TESS. Section 4 explains how changes were made to the data in the time-domain to simulate 1 year (stars in the Continuous Viewing Zone) and 27 days (stars observed for 1 sector) of TESS-like observation.

3 datasets were available to perform classification upon: the original *Kepler* data, 1 year of TESS observation, and 27 days of TESS observation. Each dataset was made up of the same 1000 stars, after their apparent magnitudes had been perturbed 100 times. Section 5 describes the detection test that was run on the individual modes of the stars in these datasets. This determined which radial solar-like modes of oscillation were detectable by *Kepler* and TESS.

Once detection probabilities were calculated for every mode, Section 6 explains how a classifier was used on the 3 datasets separately. The detection probabilities of each mode were put into three bins. Qualitatively, there were: not detected (0-50%), potentially detected (51-90%), very like to be detected (91-100%)). The global properties of the Red Giants ([log(g), π, $T_{\rm eff}$, [M/H], $I_{\rm mag}$]) were given to the classifier alongside the binned detection probabilities to train the classifier.

Once trained, the classifier was given only [log(g), π, $T_{\rm eff}$, [M/H], $I_{\rm mag}$] and required to predict the binned detection probabilities. It did this with with a 0.98% precision for the original *Kepler* data, 0.90% precision for 1 year of TESS data, and 0.81% for 27 days of TESS data.

2 THE DATASET

The data to perform Machine Learning on are the 1000 *Kepler* Red Giants from Davies et al. (in prep). In that work, the radial and quadrapole modes of these stars were fitted. Specifically, Davies et al. (in prep) provided the frequency, height, width and background of each mode in the 1000 stars. The spectroscopic parameters of these stars ($T_{\rm eff}$, log(g) , [M/H]) were available from APOKASC

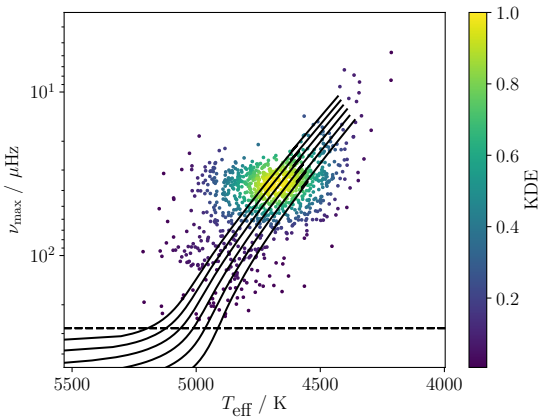

Figure 1. The 1000 *Kepler* Red Giants from Davies et al. (in prep). The colourbar shows the relative density of points with Kernel Density Estimation. The dashed line shows the *Kepler* Nyquist frequency for Full Frame Images (278Hz). The evolutionary tracks range from 0.9-1.5$M_{\odot}$. These tracks were generated using CLÉS (Scuflaire et al. 2008).

Pinsonneault et al. (2014). The apparent magnitudes are from the *Tycho*-2 catalogue Hog et al. (2000). Lastly, the parallaxes are from *Gaia* DR2 Lindegren et al. (2018). These stars are shown in Figure 1.

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). This RMS noise model takes into account 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 81,100 stars.

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 that are missing from the dataset 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. In this case, the label to predict 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);

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

The random forest regression was done in 3 stages: the algorithm was trained, tested, and used to calculate $I_{\rm mag}$ values where they are missing in the 1000-star dataset. To train and test the algorithm, the 741 *Kepler* stars with [K_p , [M/H], $T_{\rm eff}$] and known $I_{\rm mag}$ values were used. These 741 stars were split into a training dataset, and a testing dataset.

70% of the 741-star dataset was used to train the algorithm; 30% was used to test its accuracy. The results of this 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 0.06 mag. Furthermore, the standard deviation of the difference is only 0.40 mag, and the regression achieves an accuracy of 0.53. To summarise this, random forest regression predicts the $I_{\rm mag}$ values in the test dataset without introducing bias or adding a large uncertainty.

After the algorithm was trained and tested, it was used to calculate the $I_{\rm mag}$ values of the 259 stars where they are unavailable. The *Kepler* apparent magnitudes, K_p , of the stars with known and unknown $I_{\rm mag}$ values are shown in Figure 4. Similarly, the distribution of the known, and predicted (previously unknown) $I_{\rm mag}$ values is shown in Figure 5. In both Figures 4 and 5, the stars with unknown $I_{\rm mag}$ values lie toward the fainter end of the distribution. This implies that the random forest regression is making correct inferences about the $I_{\rm mag}$ values of the stars.

The majority of *Kepler* stars in the Davies et al. (in prep) dataset had measured $I_{\rm mag}$ values. Using regression, the $I_{\rm mag}$ values of the remaining stars were evaluated. After this, the lightcurves of these stars were then degraded to replicate observation by TESS.

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

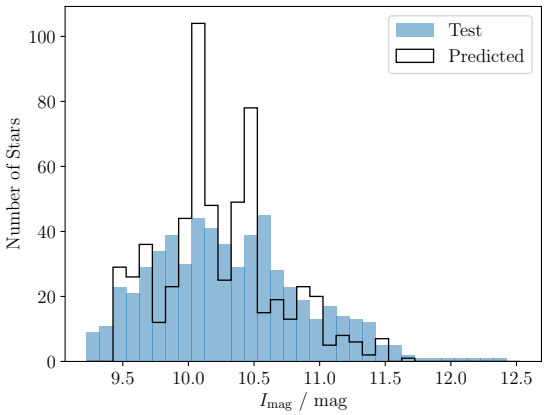

Figure 2. The $I_{\rm mag}$ distribution of the *Kepler* stars used to test the random forest regression. The 'true' values used to test the algorithm are shown in blue. The black histogram shows the distribution of values that the random forest regression predicted for those stars. A scatter plot of the differences between these distributions is shown in Figure 3.

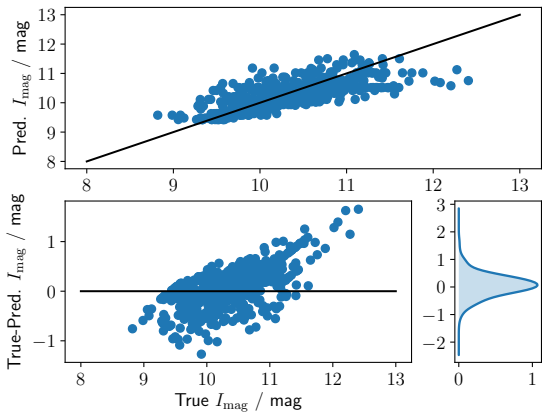

Figure 3. The true $I_{\rm mag}$ values of the *Kepler* stars use to test the algorithm, compared to their predicted values. The mean difference between the two sets of values is 0.06 mag, with a standard deviation of just 0.40 mag.

amplitude of stellar signals (i.e the signals due to stellar granulation and solar-like oscillations) (Ballot et al. 2011). Campante et al. (2016) found this correction to the oscillation intensity amplitude 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 using the model from Sullivan et al. (2015) and added to the timeseries.

These three adjustments - the length of observation, the bandpass, and the noise level - were performed in the time domain. From the original *Kepler* timeseries, the following adjustments were made:

(i) Apply a 4-σ clip to the dataset to remove spurious points.
(ii) Shorten the 4 years of timeseries data down to the reduced dataset length (either 27 days or 1 year). When reducing the length of the timeseries, take the section (27 days or 1 year) with the fewest gaps in observation out of the 4 year *Kepler* data.
(iii) Adjust the bandpass by multiplying the flux by 0.85.

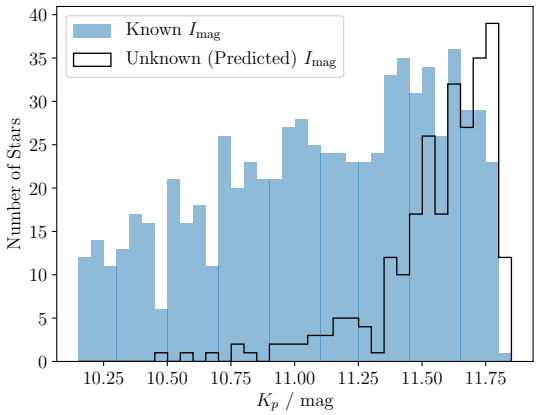

Figure 4. The K_p distribution of the *Kepler* stars with known $I_{\rm mag}$ values (741 stars) is shown in blue. The K_p values of the stars without I-band magnitudes (259 stars) are shown in black. The majority of these stars are fainter.

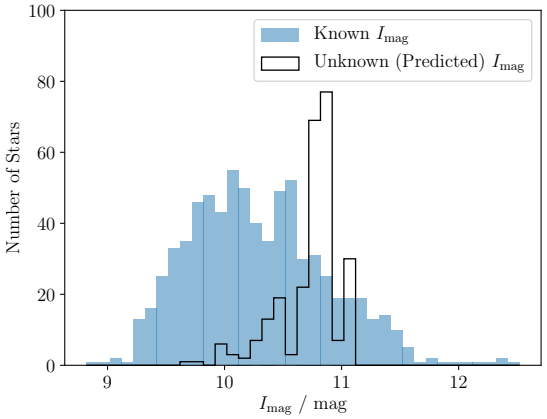

Figure 5. The $I_{\rm mag}$ distribution of the *Kepler* stars with known values (741 stars) is shown in blue. The $I_{\rm mag}$ values predicted by the regression (259 stars) are shown in black. Like in Figure 4, the predicted magnitudes are fainter than the majority of known values.

(iv) Add TESS instrumental noise to the timeseries. For each flux value in the *Kepler* timeseries ($F_{\rm Kepler}$) draw a random number from the normal distribution (N). Multiply the TESS RMS noise level ($\sigma_{\rm TESS}$, Section 2) by the random numbers. Add these to the original flux values in the *Kepler* timeseries to get the TESS flux:

$$F_{\rm TESS} = F_{\rm Kepler} + N\sigma_{\rm TESS} \quad (4)$$

(v) Compute the power spectrum from the timeseries.

The results from the original *Kepler* observation are compared to TESS-like power spectra of the same stars in Figures 6 and 7. 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.

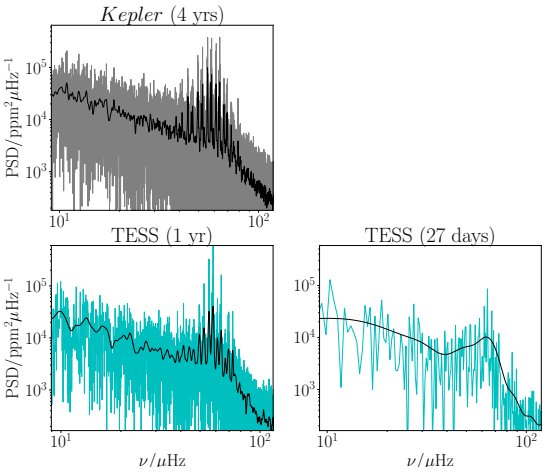

Figure 6. The Power spectra of KIC 9535399 is plotted three times with moving medians in black. The original Power Spectra is plotted in grey. The power spectra after making the data look like TESS are plotted in blue. The subplot on the bottom left shows 1 year of TESS observation (the maximum). The subplot on the bottom right column shows 27-days (the minimum).

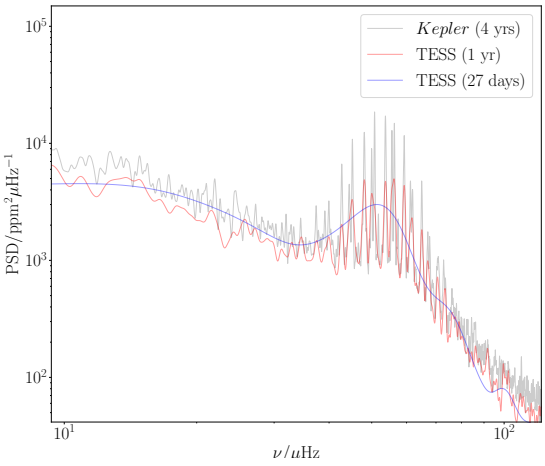

Figure 7. The Power Spectra of KIC 6768319. The original power spectra is in grey. The data were transformed into 1 year and 27 days of TESS observation. The moving median of these transformations are overplotted.

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

$$\Gamma_{\rm env} = 0.66\,\nu_{\rm max}^{0.88}, \tag{5}$$

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,

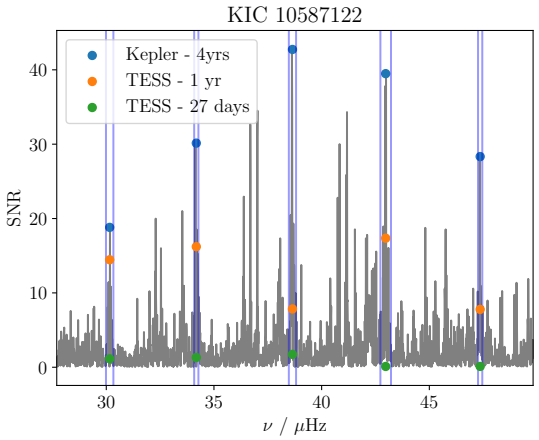

Figure 8. The SNR spectrum of KIC 10587122 after background subtraction. The SNR values of the radial modes in the star were extracted from this spectrum. The highest SNR value within the linewidth of each mode is taken to be the SNR value of that mode. The mode linewidths are shown as blue lines. The 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* .

$$\text{SNR} = P/B. \tag{6}$$

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 8.

Once all mode SNR values for the star were calculated, a detection test from Chaplin et al. (2011) was run on each mode. The detection test from Chaplin et al. (2011) 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 SNR$_{\rm thresh}$ is

$$P(\text{SNR} \geqslant \text{SNR}_{\rm thresh}) = p. \tag{7}$$

A false-alarm probability p of 5% was set; there is a 95% chance that the signal is not due to noise. Equation 7 is solved for SNR$_{\rm thresh}$ by substituting P with

$$P = \int_x^\infty \frac{e^{-x}}{\Gamma(N)} x^{N-1} dx. \tag{8}$$

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 8.

$\Gamma(N)$ is the Gamma function. The lower bound of Equation 8 is set to $x = 1 + \text{SNR}_{\rm thresh}$. The noise in the N bins is assumed to follow χ^2 $2n_{\rm bins}$ d.o.f statistics.

Once SNR$_{\rm thresh}$ is found, Equation 8 is solved again. This time it is solved for P by setting $x = (1 + \text{SNR}_{\rm thresh})/(1 + \text{SNR})$. Here, SNR is the observed signal-to-noise ratio calculated from Equation 6. This calculates the probability that the frequency spike was due to stochastic excitation in the convective envelope of the star.

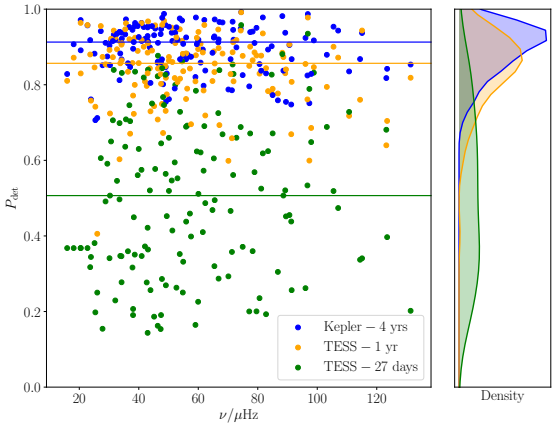

Figure 9. A plot showing the result of the detection test, after running on every mode in 20 stars. The results of the original power spectra are plotted in blue. The results of 1 year of TESS observation are in orange. 27 days of TESS observation is in green. At this short an observation, detecting individual modes will be extremely difficult.

This detection test was applied to every mode of every star in the sample. Figure 9 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[1]. 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.

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 datasets were available. In Section 4, these 81,100 datasets were treated in 3 ways: as *Kepler* targets (observed for up to 4 years), as TESS targets in the Continuous Viewing Zone (1 year of observation), and as TESS targets observed in 1 sector (for 27 days). This produced 3 sets of 81,100 datasets.

In Section 5, a mode-detectability test was performed on all 3 sets of 81,100 datasets. After this, the same classification method was performed once on each set of 81,100 datasets. Before a classifier could be used, the data needed to be prepared. That preparation is described in this Section.

Each calculated detection probability from Section 5 was put into a discrete bin (or *class*) depending on how likely the mode is to be detected. These discrete classes are given in equation 9.

$$P_{\rm det} = \begin{cases} 2 & \text{if } 1.0 \geqslant P_{\rm det} > 0.9 \ , \\ 1 & \text{if } 0.9 \geqslant P_{\rm det} > 0.5 \ , \\ 0 & \text{if } 0.5 \geqslant P_{\rm det} > 0.0 \ . \end{cases} \quad (9)$$

Using equation 9, 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.

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. There are 81,100 rows in the X-dataset.

6.2 Target selection using a classifier

The 3 sets of 81,100 datasets were prepared in Section 6.1 by assigning bins to every mode detection probability in each dataset. The first set consisted of *Kepler* targets (observed for up to 4 years) The second set comprised TESS targets that were observed in the Continuous Viewing Zone (1 year of observation). The third set contained the TESS targets that were observed in 1 sector (for 27 days). After this, the same random forest classification was performed once on each of the 3 sets of 81,100 datasets. The classifier was given 1 set of data at a time.

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

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). There are 81,100 rows in the Y-dataset. See Table 1 for the equivalent X-dataset.

The random forest classification was done in 3 stages: the data was separated, the classifier was trained, and it was tested. Firstly, the 81,100 datasets (from 811 stars) were separated into a training dataset and a testing set. 70% of the data were used to train the classifier (56,770 datasets); 30% of the stars were used to test the algorithm (24,330 datasets).

Secondly, the classifier was trained using the 56,770 datasets. The X and Y data in the training set were given to the algorithm ($\rm X_{train}$ and $\rm Y_{train}$). $\rm X_{train}$ comprises 56,770 sets of $\texttt{log}(g)$, π, $T_{\rm eff}$, $\texttt{[M/H]}$ and $I_{\rm mag}$ values. This is accompanied by $\rm Y_{train}$ with the corresponding 56,770 detection probabilities (Tables 1 and 2). The classifier trained itself by using $\rm X_{train}$ and $\rm Y_{train}$ to define distinct groups of $\texttt{log}(g)$, π, $T_{\rm eff}$ and $\texttt{[M/H]}$ in the data[2].

Thirdly, the classifier was tested with $\rm X_{test}$ and $\rm Y_{test}$. $\rm X_{test}$ comprises 24,330 sets of $\texttt{log}(g)$, π, $T_{\rm eff}$, $\texttt{[M/H]}$ and $I_{\rm mag}$ values which the classifier is given. Without having access to the corresponding $P_{\rm det}$ (1), $P_{\rm det}$ (2) and $P_{\rm det}$ (3) values in $\rm Y_{test}$, the classifier predicted a set of Y data ($\rm Y_{pred}$). 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 has 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 differences between the 'true' detection probabilities ($\rm Y_{test}$) and the values predicted by the classifier ($\rm Y_{pred}$) were calculated. This was done for each feature ($P_{\rm det}$ (1), $P_{\rm det}$ (2) and $P_{\rm det}$ (3)) separately. The mean of these differences was calculated, and weighted by the number of true-positive values in each feature.

This weighted precision $\mathcal{P}$ is given as

$$\mathcal{P} = t_{\rm p}/(t_{\rm p} + f_{\rm p}), \tag{10}$$

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 $\rm Y_{test}$;

$$H_{\rm loss}(Y_{\rm test}, Y_{\rm pred}) = \frac{1}{n_{\rm classes}} \sum_{j=0}^{n_{\rm classes}-1} 1(Y_{\rm pred} \neq Y_{\rm test}) \, . \tag{11}$$

[2] https://docs.marklogic.com/
[3] http://scikit-learn.org

Satellite	$T_{\rm obs}$	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 [$P_{\rm det}$ (1), $P_{\rm det}$ (2), $P_{\rm det}$ (3)].

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 gauged. 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 9, 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. The 3 classes and $P_{\rm det}$ ranges given in equation 9 gave the best predictions for the 3 sets of 81,100 datasets (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 TESS data are very good; the classifier was able to replicate the mode detection predictions of the stars robustly. For the 27-day TESS dataset, the classifier was able to correctly predict the $P_{\rm det}$ values of 81% of the modes. This is because with these briefly-observed TESS targets, the data is very susceptible to the realization noise of the observation. While an individual mode may be detectable in one 27-day observation, the high realisation noise may render it undetectable in 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 is that it returns the 'feature importance' of each label in the X-data.

As Table 1 shows, the X-data labels (or features) that are given to the classifier are $\texttt{[log}(g)\texttt{,}$ π, $T_{\rm eff}$, $\texttt{[M/H],}$ $I_{\rm mag}\texttt{]}$. Some of these X-data labels are more important than others when predicting the detection probability of solar-like modes $\texttt{[}P_{\rm det}\texttt{(1),}$ $P_{\rm det}\texttt{(2),}$ $P_{\rm det}\texttt{(3)]}$. The feature importance of an X-data label is a measure of how informative that label is when predicting $\texttt{[}P_{\rm det}\texttt{(1),}$ $P_{\rm det}\texttt{(2),}$ $P_{\rm det}\texttt{(3)]}$.

The feature importance of an X-data label is calculated as the number of splits in the decision trees that include that label, divided by the total number of splits in all of the decision trees. For example, if parallax is used in 10% of all the splits made across all of the decision trees, then parallax will have a feature importance of 0.1. The feature importance of all 5 X-data labels sums to 1.

The feature importances of the X-data labels are given in Figure 10. The feature with the highest influence on mode detectability is apparent stellar magnitude $I_{\rm mag}$. This is not surprising: fainter stars have higher shot-noise levels. This will reduce the signal-to-noise ratio (equation 6) of the modes in those stars, making them less likely to be detected.

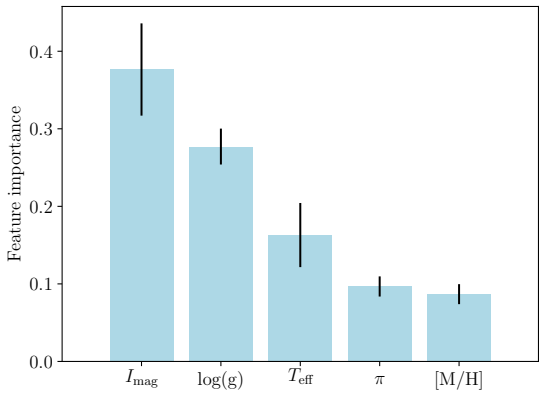

Figure 10. The feature importance of the 5 X-data labels.

After the apparent stellar magnitude, surface gravity $\log(g)$ also has a heavy influence on P_{det} value. This can be explained in terms of the oscillation amplitude of stars. Kjeldsen & Bedding (1995) showed that the oscillation amplitude observed in intensity is proportional to the bolometric luminosity, mass and effective temperature of the star;

$$A_{osc} \propto \frac{L}{MT_{eff}^{0.5}}. \quad (12)$$

By making substitutions into equation 12, the oscillation amplitude can be shown to be proportional to $\log(g)$ and T_{eff} alone. Using $\log(g) \propto M/R^2$ and $L \propto R^2 T_{eff}{}^4$, equation 12 becomes

$$A_{osc} \propto \frac{T_{eff}{}^{3.5}}{\log(g)}. \quad (13)$$

As a star evolves 'up' the Hertzsprung-Russell diagram, its surface gravity decreases while its effective temperature stays roughly the same (see the evolutionary tracks in Figure 1). 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.

Equation 12 shows that oscillation amplitude has a larger dependence on effective temperature than on surface gravity. T_{eff} might therefore be expected to have a higher feature importance than $\log(g)$ when predicting $[P_{det}(1),\ P_{det}(2),\ P_{det}(3)]$. Despite this, Figure 10 shows that T_{eff} is less informative than $\log(g)$ when predicting detection probability. This is because T_{eff} varies very little for Red Giant stars, see Figure 1.

If an X-data label has a low feature importance, this leads to two conclusions; either the label is *irrelevant*, or it is *redundant*. An irrelevant label is completely unrelated to the features that are being predicted. A redundant label is relevant to the feature being predicted, but is very similar to one or more other X-data labels in the training dataset. For example, including both I_{mag} and V_{mag} X-data labels would result in one magnitude being disregarded by the classifier, and having a low feature importance score. Although Figure 10 shows that apparent magnitude is an important feature when predicting detection probability, including both I_{mag} and V_{mag} would not give the classifier more information to predict detection probability compared to only including I_{mag} . V_{mag} would be a redundant label.

Parallax π has a relatively low feature importance when predicting detection probability. Parallax is proportional to the inverse distance to the star, which is related to the apparent magnitude of the star. This makes parallax a partly redundant label: when an apparent magnitude is also included in the list of X-data labels, adding parallax provides less extra information than if I_{mag} were not included.

As well as apparent magnitude, parallax is also related to luminosity (for example, see Torres et al. (2010)). Luminosity is proportional to oscillation amplitude, so including parallax does provide a small amount of extra information to the classifier.

Metallicity [M/H] is the least informative label when predicting P_{det}. Oscillation amplitude does not have a large dependence on metallicity; metallicity is less relevant than the other X-data labels when predicting detection probability.

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 $\log(g)$.

Evolution	Satellite	T_{obs}	Precision	Hamming loss
RGB	*Kepler*	4 years	0.99	0.01
RGB	TESS	1 year	0.88	0.11
RGB	TESS	27 days	0.82	0.18
RC	*Kepler*	4 years	0.98	0.02
RC	TESS	1 year	0.92	0.07
RC	TESS	27 days	0.82	0.19
2CL	*Kepler*	4 years	0.98	0.02
2CL	TESS	1 year	0.85	0.13
2CL	TESS	27 days	0.75	0.26

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)].

6.4 Comparing results between different evolutionary states

So far in this paper, 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.

If a Red Giant Branch is massive enough, it will undergo the Helium flash and become a Red Clump star. If a star is more massive than $\gtrapprox 1.8 M_{\odot}$, it will instead become a Secondary Clump star. The evolutionary states of the stars in the sample were taken from Elsworth et al. (2017).

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 $P_{\rm det}$.

7 CONCLUSION

The solar-like oscillations of 1000 *Kepler* Red Giant stars from Davies et al. (in prep) were used to show that asteroseismic target selection can be done using a classifier. A classifier was used here 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 classifier can also be used to understand target selection bias in previous missions, such as K2 and CoRoT.

Firstly, the number of datasets was increased by perturbing the stellar magnitudes 100 times for each star. These perturbed magnitudes were drawn from a PDF of the *Kepler*n and TESS shot noise models (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 6).

A detection test was then run on the SNR values at every mode frequency (Section 5). This gave a detection probability $P_{\rm det}$ between 0.0 and 1.0 for every mode. In order to prepare the detection probabilities before classification, each continuous $P_{\rm det}$ value was assigned a discrete class ([0,1 or 2]; equation 9).

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 $[\log(g),\ \pi,\ T_{\rm eff}\ ,\ [{\rm M/H}],\ I_{\rm mag}\ ]$ 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 $P_{\rm det}$ values of 3 radial modes centred around $\nu_{\rm max}$ 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 of 0.98 across the 3 $P_{\rm det}$ 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. When the different evolutionary states were kept together, the predicted mode detections scored a weighted precision of 0.90 across the 3 $P_{\rm det}$ features (Table 3). 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. While individual modes may be detectable in one observation, a different 27-day observation (and noise realisation) may render the modes undetectable. 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

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

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

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

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

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

Scuflaire R., ThÃľado S., MontalbÃąn J., Miglio A., Bourge P.-O., Godart M., Thoul A., Noels A., 2008, Astrophysics and Space Science, 316, 83

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

Torres G., Andersen J., GimÃľnez A., 2010, Astronomy and Astrophysics Review, 18, 67

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.