Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **455,** 626–633 (2015) doi:10.1093/mnras/stv2333

SIDRA: a blind algorithm for signal detection in photometric surveys

D. Mislis,[★] E. Bachelet, K. A. Alsubai, D. M. Bramich and N. Parley

Qatar Environment and Energy Research Institute (QEERI), HBKU, Qatar Foundation, PO Box 5825, Doha, Qatar

Accepted 2015 October 6. Received 2015 October 5; in original form 2015 April 13

ABSTRACT

We present the Signal Detection using Random-Forest Algorithm (SIDRA). SIDRA is a detection and classification algorithm based on the Machine Learning technique (Random Forest). The goal of this paper is to show the power of SIDRA for quick and accurate signal detection and classification. We first diagnose the power of the method with simulated light curves and try it on a subset of the *Kepler* space mission catalogue. We use five classes of simulated light curves (CONSTANT, TRANSIT, VARIABLE, MLENS and EB for constant light curves, transiting exoplanet, variable, microlensing events and eclipsing binaries, respectively) to analyse the power of the method. The algorithm uses four *features* in order to classify the light curves. The training sample contains 5000 light curves (1000 from each class) and 50 000 random light curves for testing. The total SIDRA success ratio is ≥90 per cent. Furthermore, the success ratio reaches 95–100 per cent for the CONSTANT, VARIABLE, EB and MLENS classes and 92 per cent for the TRANSIT class with a decision probability of 60 per cent. Because the TRANSIT class is the one which fails the most, we run a simultaneous fit using SIDRA and a Box Least Square (BLS)-based algorithm for searching for transiting exoplanets. As a result, our algorithm detects 7.5 per cent more planets than a classic BLS algorithm, with better results for lower signal-to-noise light curves. SIDRA succeeds to catch 98 per cent of the planet candidates in the *Kepler* sample and fails for 7 per cent of the false alarms subset. SIDRA promises to be useful for developing a detection algorithm and/or classifier for large photometric surveys such as *TESS* and *PLATO* exoplanet future space missions.

Key words: techniques: photometric – planets and satellites: detection – planets and satellites: fundamental parameters – planetary systems.

1 INTRODUCTION

The recent development of wide-field photometric surveys opens up a new field of astrophysics. The deployment of both ground-based (SuperWASP, HAT, QES) and space-based (*CoRoT*, *Kepler*) surveys increases dramatically our knowledge about transiting planets. Indeed, the huge amount of collected data leads to a real problem of identifying targets. The OGLE survey, for example, observes more than 300 million stars in the Galactic bulge each night, leading to a sorting problem. This kind of problem is a known as a Big Data problem.

Several methods are used to tackle the Big Data problem, and the Machine Learning algorithm is one of them. The Random Tree/Forest algorithm was described in 2001 by L. Breiman (Breiman 2001) as part of Artificial Intelligence and Machine Learning general algorithms. Some teams have already used Machine Learning algorithms for astronomical projects, especially for the large amount of data from the *Kepler* mission (Hogg et al. 2013). Machine Learning object detection and classification for automated classification of active stars and galaxies is described in Li, Zhang & Zhao (2008), using the k-Nearest Neighbours method. Recently, Masci et al. (2014) published an algorithm based on Random Forest for automatic classification of variable stars using the *Wide-field Infrared Survey Explorer* data with a success ratio from 87.8 to 96.2 per cent. In the same year, OGLE detected a supernova Type Ia event, using real-time detection and Machine Learning automatic classification (Wyrzykowski et al. 2014). Furthermore, a Random Forest algorithm is used by McCauliff et al. (2014) to identify false positives in the *Kepler* mission data.

The exoplanet microlensing surveys such as OGLE and MOA are facing a challenge with real-time photometry and lens detection (Bond et al. 2001). Microlensing detections must be observed by as many teams as possible in order to have a complete phase coverage of the phenomenon. This introduces a need for fast event detection on a huge amount of light curves.

In 2002, Kovács, Zucker & Mazeh (2002) published the Box Least Square (BLS) algorithm for transiting exoplanet detection and since then, there are many different versions of BLS (Foreman-Mackey et al. 2015). BLS is a very successful algorithm for

[★] E-mail: dmislis@qf.org.qa

Table 1. Input parameters and limits for our simulated light-curves sample. For each value we use uniform random distribution.

rms (ϵ)	0.01–0.05
Period (P)	1–15 d
Spectra type	F0–M5
Planetary radius ($R_{\rm P}$)	0.7–2.0$R_{\rm J}$
Transit inclination (i)	$i_{\rm min}$–90 deg
Observing window (t)	30 d
Time resolution	30 min

almost all of the transiting surveys such as *Kepler* (Boruki et al. 2010), *CoRoT* (Moutou et al. 2007) – from space – and Super-WASP (Cameron et al. 2006), HATNet (Bakos et al. 2011), QES (Alsubai et al. 2013) etc. – from the ground. In principal, in order to detect a transiting signal in a light curve using a BLS algorithm, we have to fit the orbital period of the planet, the centre of the eclipse $T_{\rm c}$, the duration of the transit and the depth of the transit (Bordé et al. 2007; Bonomo et al. 2012; Cabrera et al. 2012).

In this paper, we study a very different approach for signal detection and classification for transiting exoplanets, variable stars and microlensing events by changing the philosophy of signal detection from fitting to blind search using Machine Learning techniques.

2 DESCRIPTION OF THE METHOD AND SIMULATED LIGHT CURVES

2.1 Light-curve simulations

In this paper, we used simulated light curves for the training and testing sample in order to perform various tests. We focused on five typical light-curve types which can be expected in photometric surveys. These are constant stars (called hereafter CONSTANT), the exoplanet transiting light curves (TRANSIT), the variable stars (VARIABLE), the eclipsing binaries (EB) and the microlensing light curves (MLENS). Each light curve is described by a normalized flux as a function of time. We added noise to each light curve with various precisions. The rms is selected from a uniform distribution between 1 and 5 per cent. We used a 30-d observing window, with a 30 min sample to simulate the light curves. Some of these types of light curves, such as the TRANSIT sample, require stellar physical properties (stellar mass, stellar radius, effective temperature) given by Kaler (1998). We select main-sequence host stars randomly, using a uniform distribution from F0 to M5 spectral type. This range was adopted because it roughly represents 90 per cent of the total stars in the sky (Robin et al. 2004). Table 1 summarizes all the input parameters and ranges we used for the simulated light curves. Fig. 1 shows typical simulated light curves from each class. A more detailed description for each light-curve class is given below.

2.1.1 CONSTANT

This subset of light curves is the most simple, and we can easily create it using pure white noise. The flux $f_{\rm C}$ of a constant light curve is given by

$$f_{\rm C}(t) = 1 + \epsilon, \tag{1}$$

where ϵ is a random variable set by a normal distribution N(0, rms). The rms is randomly selected to be in the range of $0.01 \le {\rm rms} \le 0.05$.

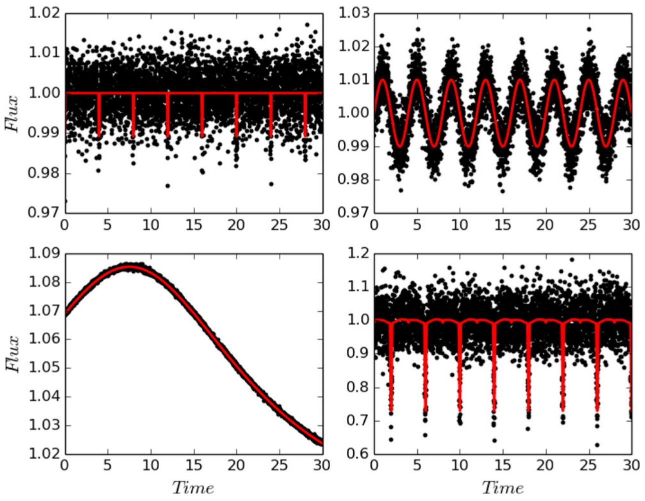

Figure 1. Four random light curves (TRANSIT, VARIABLE, MLENS and EB) and their model (red solid line).

2.1.2 TRANSIT

The host stars' spectral type and the physical properties are selected randomly from the main-sequence data set as we describe in Section 2.1. The planet radius is also chosen randomly to be in the range 0.7–2$R_{\rm J}$ and the period was chosen between 1 and 15 d. The inclination angle i was chosen in the range $i_{\rm min} \le i \le 90$ deg to ensure a transit. The $i_{\rm min}$ is the minimum transit angle described by

$$\cos i_{\rm min} = \frac{R_\star + R_{\rm P}}{\alpha}, \tag{2}$$

where $R_\star$ is the star radius, $R_{\rm P}$ the planetary radius and α the semi-major axis. Note that we choose a null eccentricity. The transit model is produced using a quadratic limb-darkening law and the adopted flux is given by

$$f_{\rm T}(t) = {\rm Pal}(t, P, \delta, i, d) + \epsilon, \tag{3}$$

where Pal() is the Pál (2008) transiting analytical model, P is the period, δ is the transit depth and d is the duration of a transit. The limb-darkening coefficients are given by Claret (2004).

2.1.3 VARIABLES

For this subset, we selected only stellar types crossing the main sequence and the instability strip of the Hertzsprung–Russel diagram. This leads to select only F-type stars. We modelled the light curve by using a pure sinusoidal signal, with a flux amplitude $0.01 \le A_{\rm V} \le 0.1$ and period P, between 1 and 15 d. This way we mainly modelled some common types of variable stars such as RR Lyrae, δ-Scuti etc. Equation (4) gives the flux of our variable subset light curves:

$$f_{\rm V}(t) = 1 + A_{\rm V} \cdot \sin(2\pi t/P) + \epsilon. \tag{4}$$

2.1.4 Eclipsing binaries

The eclipsing binaries subset light curves were modelled using the same stellar main-sequence characteristics (Section 2.1). The selected period is again between 1 and 15 d. In order to simulate a full eclipsing binary light curve (primary/secondary eclipse, ellipsoidal variations etc.) we used the PHOEBE eclipsing binary analytical model (Prša 2005). PHOEBE requires stellar characteristics for both of the stars, such as stellar temperatures, masses and radii plus the orbital period of the system. We do not describe here the full equation package but the light-curve flux $f_{\rm EB}$ we used is given by equation (5):

$$f_{\rm EB}(t) = \text{PHOEBE}(t, M_i, R_i, T_i, P) + \epsilon, \quad (5)$$

where PHOEBE() is the eclipsing binaries model, M_i, R_i and T_i are the stellar mass, radius and temperature, respectively, for the primary ($i = 1$) and the secondary ($i = 2$) star. Finally, P is the orbital period.

2.1.5 Microlensing

We produce microlensing light curves as though they have been observed by a survey such as MOA (Bond et al. 2001) or OGLE (Udalski et al. 1992). We select a uniform random value for the time of maximum $t_{\rm o}$ between 0 and 30 d. We also select $U_{\rm o}$, the minimum impact parameter, from a uniform distribution (Alcock et al. 2000; Sumi et al. 2011). We select the Einstein ring crossing time $t_{\rm E}$ from a normal distribution with a mean of 20 d and a standard deviation of 5 d, which is a rough approximation of the true $t_{\rm E}$ distribution (Sumi et al. 2011). Finally, we select a uniform distribution for the source flux $f_{\rm s}$ and the blend flux $f_{\rm b}$ in the range of 1–10. The adopted flux is

$$f_M(t) = \frac{f_{\rm s} \cdot A(t) + f_{\rm b}}{f_{\rm s} + f_{\rm b}} + \epsilon, \quad (6)$$

where $A(t)$ is the microlensing magnification (Paczynski 1986).

2.2 Random forest basics

In the Machine Learning domain, the key is to give informative parameters to the algorithm that describe the problem in hand. These parameters, called features, must reflect the intrinsic properties of the different classes. Using these features as inputs, the Random Forest is a three-step algorithm, as a typical Machine Learning procedure suggests (train–test–predict).

(i) The first step is the training part of the algorithm. The Random Forest uses the features of each vector of the training sample to build $N_{\rm tree}$ decision trees which are tuned to fit the output classes (train-step).

(ii) After the training process, it is recommended to characterize the performance of the algorithm by using an exercise sample (test-step).

(iii) If the user is satisfied with the accuracy of test-step, then the Random Forest can be used for any feature of the sample (predict-step).

To help in the customization of the Random Forest algorithm, we used various tools. The *feature importance* vector gives the relative importance of each feature to produce the most accurate estimator. The *confusion matrix* shows in a simple way how well the algorithm performs. Its diagonal values are equal to the *success ratio* for each of the classes. Also, the $i \neq j$ element of the *confusion matrix* give the false positive/negative rates (Masci et al. 2014).

2.3 The statistical method

SIDRA, basically follows three simple rules.

(i) Features must be as general as possible. SIDRA is able to compute them in a fully blind way for all kinds of light curves.

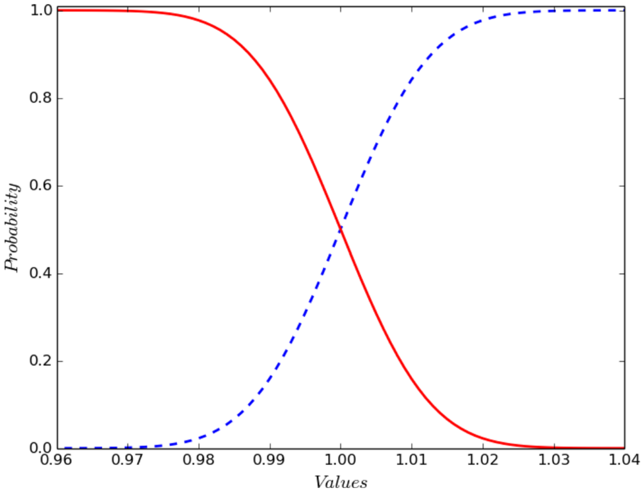

Figure 2. Two different probability functions. A normal Gaussian CDF (blue dashed line) and the inversed Gaussian CDF (red solid line).

(ii) Features' extraction must also be as fast as possible, in order to make the algorithm useful for large and/or real-time surveys (*TESS*, *PLATO*, OGLE, ATLAS etc).

(iii) Features must show very weak correlation with each other. There is no additional information for highly correlated features.

To respect the first and second conditions, we make the choice to derive our features without any model fits and we calculate only the *statistical information* straight from the light curve. We use four features for SIDRA.

(i) The skewness S: this is a measure of the asymmetry of a distribution, defined as the third standardized moment –

$$S = \frac{1}{n} \sum_{i=1}^{n} \frac{(x_i - \bar{x})^3}{\sigma^3}, \quad (7)$$

where $\bar{x}$ is the mean, σ the standard deviation and n the total number of observations.

(ii) The kurtosis K: this is a measure of the flatness of a distribution defined as the fourth standardized moment –

$$K = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^4}{\sigma^4}. \quad (8)$$

(iii) The autocorrelation integral $A_{\rm I}$: the autocorrelation integral is the sum of the autocorrelation values for all possible delays τ. The autocorrelation versus delay (τ) diagram gives information about periodical patterns of the light curve. For SIDRA, we explore the full observation window and measure the integral given by the autocorrelation vector

$$A_{\rm I} = \left| \sum_{\tau=1}^{n} \left(\frac{1}{(n-\tau) \cdot {\rm rms}^2} \sum_{i=1}^{n=\tau} (x_i - \bar{x})(x_{{\rm i}+\tau} - \bar{x}) \right) \right|. \quad (9)$$

(iv) The modified information entropy ($E_{\rm S}$) – or Shannon Entropy (Shannon & Weaver 1949): for each class of light curve, we assume a normal distribution. This is true only for a CONSTANT light curve, but still there is more information for all the other light-curve types too. Thus, each x_i has a probability based on the Cumulative Distribution Function (CDF). Based on the nature of the survey (exoplanets, variables, microlensing), we can use different CDFs (normal or inversed Gaussian CDF – Fig. 2) in order to cover different light-curve cases. For our current SIDRA version we used the normal and the inversed Gaussian CDF (blue and red line –

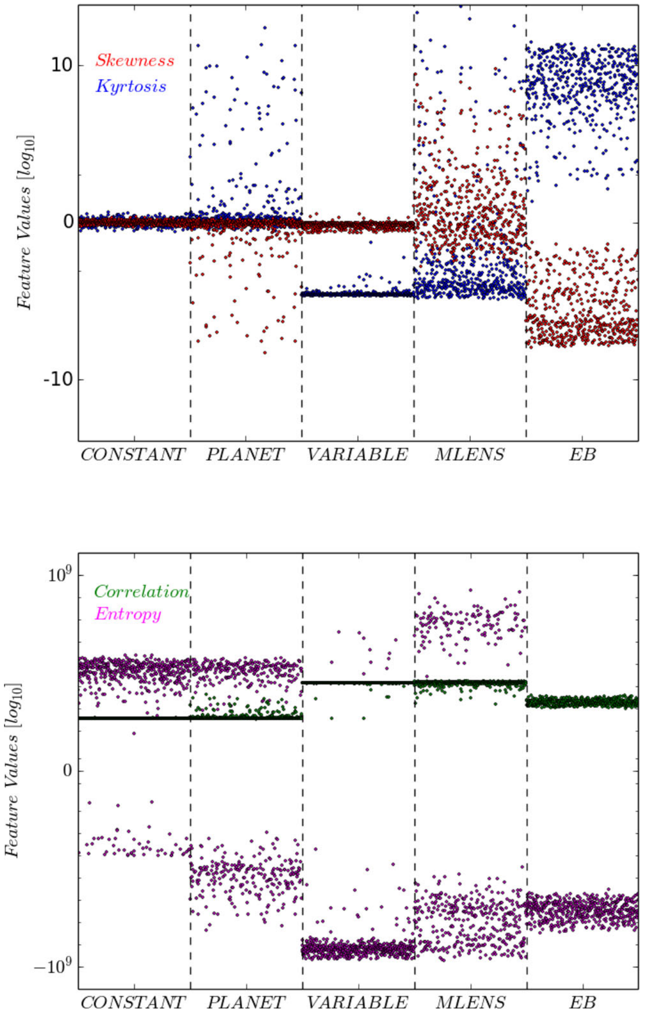

Figure 3. Different features for all classes. Top: the skewness and kurtosis (red and blue, respectively). Bottom : autocorrelation and entropy features (green and magenta, respectively).

Fig. 2) in combination. The probability $P(x_i)$ of each data point x_i is given by equation (10):

$$P(x_i) = \frac{1}{2} \cdot \left(1 + \frac{2}{\sqrt{\pi}} \int_0^{\frac{x_i - \bar{x}}{\mathrm{rms} \cdot \sqrt{2}}} \mathrm{e}^{-t^2} \mathrm{d}t \right) , \tag{10}$$

and the total entropy E_S of each light curve is given by equation (11):

$$E_\mathrm{S}(x) = -\sum_{i=1}^{i=n} \left(\int_{\delta_1}^{\delta_2} \log_2(P(\delta))\mathrm{d}\delta \right) , \tag{11}$$

where $\delta_{1,\,2} = x_i \pm \sigma_i$ and σ_i is the error of the point x_i of the light curve.

Our final E_S is calculated by adding two values of E_S calculated by the normal and inverse Gaussian CDF.

The features we have chosen show weak correlation in the parameter space. Fig. 3 shows the correlation matrix between all features and classes. VARIABLE, EB and MLENS classes are very well determined. On the other hand there is a confusion between CONSTANT and TRANSIT classes. Because of the low signal-to-noise (S/N) ratio of some light curves, it is impossible to distinguish between constant and transits (pure noise light curves). Fig. 3 (bottom) explains the results of the *confusion matrix* (Fig. 5).

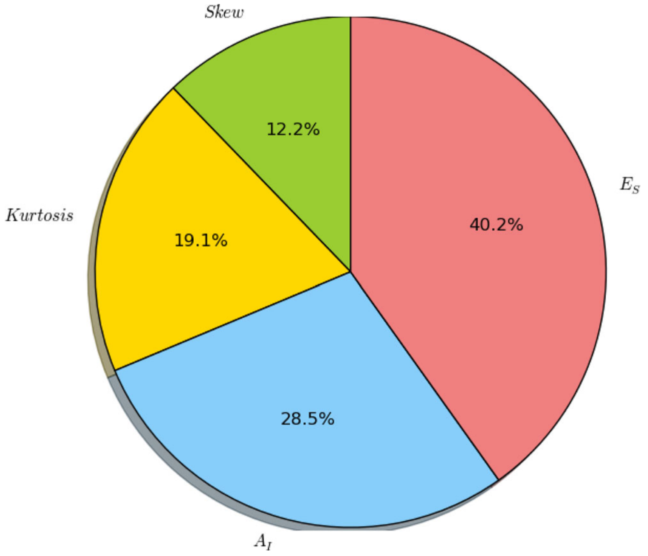

Figure 4. The feature importance statistics. E_S value is the most important (40.2 per cent).

It is clear from Fig. 3 (top) that in some cases the Random Forest decision is very obvious because the classes are very well separated, suggesting that maybe we do not need a Random Forest algorithm. On the other hand, Random Forest becomes important to distinguish objects where their features are mixed (Fig. 3 – bottom). Also, in this paper we show only some cases. The input classes could be modified by any team, or even increased by adding more different light curves, making the problem much more complicated (distinguish between supernova – microlensing light curves or different types of variables for example).

3 PERFORMANCE

3.1 General

First we create a ‘TRAIN’ sample using the five classes of light curves described in Section 2.1. We use 1000 light curves for each class (5000 in total) and we calculate the S, K, A_I and E_S features of each one. We also add a flag (CONSTANT, TRANSIT, VARIABLE, EB, MLENS) for each algorithm decision per class. By training SIDRA we found that the fitting score of the ‘TRAIN’ sample is 90 per cent using 100 trees. That means from the 5000 light curves, SIDRA could successfully distinguish 4500 of them. The majority of the remaining 10 per cent, which SIDRA fails, comes from CONSTANT and TRANSIT due to low S/N ratio, as we have described previously.

It is very important for all features to have a good statistical weight in the procedure. The importance of each feature is high enough to be included in the algorithm. After the training procedure, the importance of each feature is shown in Fig. 4.

The most important feature is E_S with 40.2 per cent, then A_I with 28.5 per cent and skewness and kurtosis with 12.2 and 19.1 per cent, respectively. The A_I feature contains high values for high-amplitude light curves such as microlensing and/or variables. On the other hand, skewness and kurtosis include information on the light curve shape, which is different for different classes. Finally, E_S shows

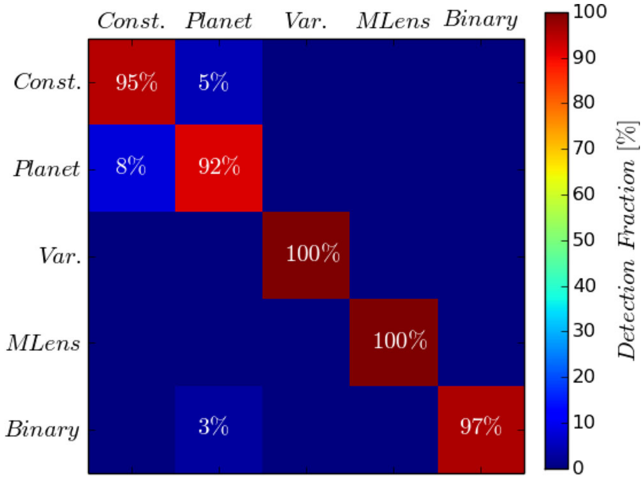

Figure 5. The confusion matrix for decision probability of 0.6, where rows and columns refer to input and output light-curve flag, respectively. The colour bar refers to the success ratio.

values around zero for variables, and constant, with high positive values for microlensing and negative values for planets and binaries.

We applied SIDRA to 10 000 light curves from all classes as a blind test (50 000 light curves in total). We create a *confusion matrix* using these results. In principal, the algorithm collects all decisions from all different trees. The final decision is made by maximizing the probability from all of the different trees. In the worst case, the minimum decision probability is 0.2 (because we have five classes – 1/5), for a five class Random Forest such as SIDRA.

It is obvious that the small decision probability decreases the success ratio of the algorithm because we have to deal with a flip-coin decision. We force SIDRA to take more certain decisions. For our example, we used a decision probability equal to 0.6. The *confusion matrix* shows the results of this test (Fig. 5). The success ratio for each class is 100 per cent for microlensing and variable stars, 97 per cent for eclipsing binaries (3 per cent planet false alarm), 95 per cent constant (5 per cent planet false alarm) and 92 per cent for transits (8 per cent constant false alarms). If we increase the decision probability (from 0.2 to 0.6) some light curves in the range of 0.2–0.6 are rejected. At 0.6 decision probability the algorithm rejects 5 per cent of the total sample.

It is clear that the success ratio of the algorithm is a function of the decision probability cut and there are no 'golden' fixed numbers for each survey. On the other hand, each survey should define these numbers for their own goals, depending on their targets and features. As an example we can say that it is extremely rare for a transit survey to detect a microlensing event. Most of the transit surveys (if not all) avoid the fields in the Galactic plane. In these fields the probability to detect a microlensing event is close to zero. On the other hand microlensing surveys do not search for transiting planets because of the magnitude range and faint target stars of the field.

We plan to give a more detailed analysis for the decision probability based on different algorithms (such as Bayesian, Dempster–Shafer theory and/or Fuzzy Logic) in a future paper.

3.2 A closer look at planets

From the tests in Section 3.1, the most confusing classes are CONSTANT and TRANSIT. Fig. 5 suggests that 92 per cent of the transiting light curves can be resolved by SIDRA, but this is not totally true. In Table 1 we select all the host stars between F0 and M5, planetary radius 0.7–2R_J and noise from 1 to 5 per cent. That means that for 30 per cent of our stellar sample (F0–G0), the transit depth range is from 0.008 to 2.5 per cent for the worst and best case, respectively. Most of these planets do not show any signal in the light curve with the noise properties we used and it is impossible to be detected. The real question is how many 'detectable' transiting planets SIDRA could flag.

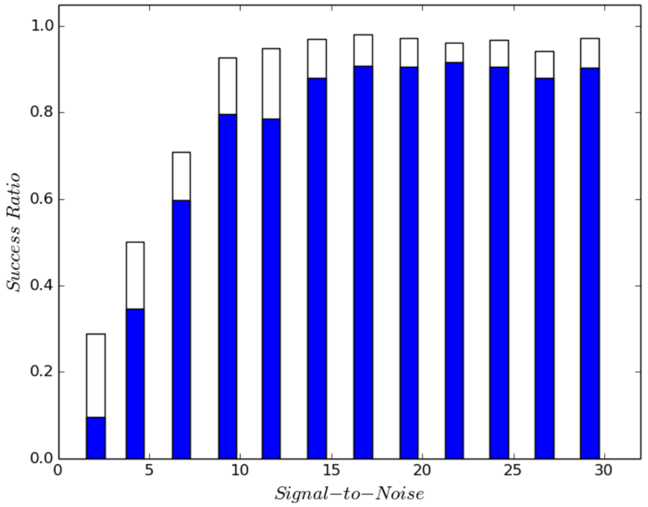

Figure 6. BLS (blue bars) and SIDRA (white bars) results after simultaneously searching for different S/N values.

In order to judge the algorithm on the transiting light curves sample, we compare SIDRA with a BLS-type algorithm, such as Kovács et al. (2002). We run BLS and SIDRA simultaneously on the same data set. For BLS we select a signal detection threshold at the 1σ level. Even if 1σ is not realistic for a real world survey (we expect signals above 2σ), this threshold is generous for BLS. If we increase the detection threshold, of course we expect much less planets. We compare with SIDRA 0.5 decision probability threshold (Fig. 6).

Both of the algorithms found approximately the same amount of planets. SIDRA detected 85.4 per cent of the sample and BLS 77.9 per cent of the sample (7.5 per cent less planets than SIDRA) even with 1σ detection threshold. Also, Fig. 6 shows that SIDRA is more sensitive than BLS for low S/N light curves.

3.3 Real data example – *Kepler* mission

The next task was to use our algorithm on real data in order to check its success. This section is only a small example with limited amount of data and tests. We plan to present a full *Kepler* mission data analysis using SIDRA in a future publication. For our tests we focus only on the exoplanets group of light curves using *Kepler* public light curves available from the *Kepler* archive hosted by the Multimission Archive at STScI.[1] The observations comprised only from the long cadence.

In order to use these data from a space mission, we modified our training sample. We did not train for variables, microlensing events or eclipsing binaries.

We have used *Kepler* Q1–Q6 Kepler Object of Interest (KOI) data set. We used 2000 light curves flagged as transit candidates (PLANET-SET; Batalha et al. 2013; Mullally et al. 2015) and 2000

[1] http://archive.stsci.edu

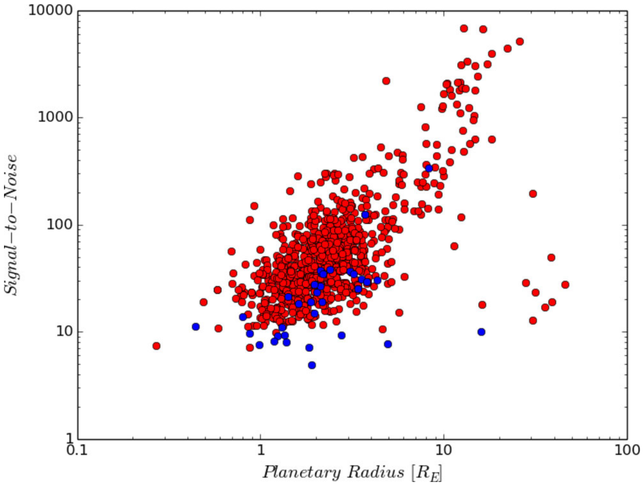

Figure 7. The *Kepler* sample we used for our test. From 0.5 to 50R_E and 5–5000 S/N. The S/N has been calculated in a 400.0-d orbital period window. Plot shows also the SIDRA successful and unsuccessful detections (red and blue dots, respectively).

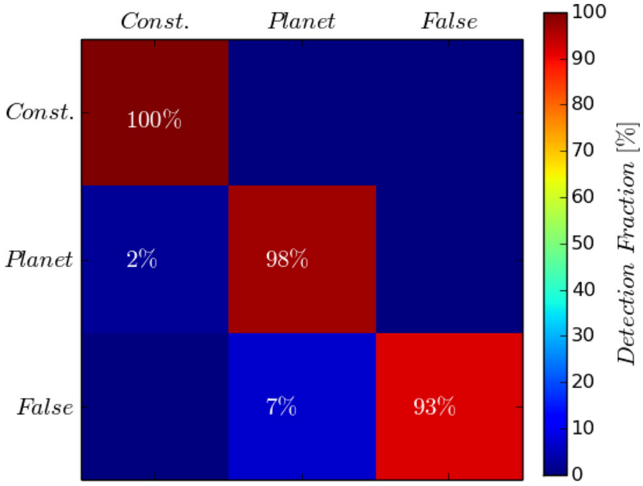

Figure 8. The confusion matrix of the *Kepler* data test, where rows and columns refer to input and output light-curve flag, respectively. The colour bar refers to the success ratio. Decision probability is 0.5.

light curves from the Q1–Q6 data set flagged as non-transiting exoplanets (CONSTANT-SET). We select our sample randomly, which means that our sample includes both large and small transiting light curves. Fig. 7 shows a planetary radius versus S/N ratio of the *Kepler* sample we used.

Furthermore we create another data set with 1000 light curves flagged as FALSE-SET. These light curves are included in the KOI *Kepler* catalogue but they are not real planets. We select false alarm light curves which belong to the CONSTANT-SET. Finally, we have to mention that we were using ∼400 d of Q1–Q6 KOI data set.

We train the forest using two classes (CONSTANT and TRANSIT) as described in Section 2.1, but we run it for all three data sets. We used rms range from 0.001 to 0.01, period range 1–200 d and planetary radius of 1–5R_E. The total number of training light curves was 1000 per class. Fig. 8 shows the results of our test.

The success ratio of real planets is 98 per cent. The constant success ratio is 100 per cent and the False Alarm success ratio is 93 per cent. The False Alarm ratio is quite important. SIDRA classifies only the 7 per cent of the False Alarm light-curves as planets, which appear in the KOI *Kepler* list. On the other hand SIDRA seems to miss 2 per cent of real planets. In order to detect planets with higher period and/or smaller S/N, we need many more data than 400 d. Also, we did not include any ‘exotic’ light curves in our training sample (Section 3.4). A more accurate analysis is required in a future paper including many more classes other than Constant and Planet data sets. The decision probability of 0.5 contains the 90 per cent of the sample.

3.4 Exotic light curves

The *Kepler* mission data have shown how difficult it is to detect transiting exoplanets around a star with high variability. These kind of light curves are the most important for space missions because with such high photometric accuracy, most of the stars show some kind of real variability. BLS-like algorithms fail to detect these kind of transits because of the algorithm design. BLS assumes that the out-of-transit mean value of all transits is 1 (or 0). This is not true in a variable star light cure with transit. The signal of the variable star dominates the light curve with very strong primary and harmonic periods. Almost all the combined transiting light curves need special analysis. For a pure blind detection method it is a very difficult problem for any algorithm, including Machine Learning.

SIDRA is able to ‘solve’ the problem with a combined analysis method similar to other BLS-like techniques. We simulate multi-period variable plus transiting exoplanet signals using *Kepler* accuracy specifications. Fig. 9 (top) shows an example of our simulated light curve.

The strategy is simple. We first run SIDRA on the raw light curve. The algorithm classifies the light curve as a variable with probability 98 per cent. Once we detect a variable signal, we use Lomb–Scargle, FFT or binned polynomial fits in order to remove strong periodicities (Fig. 9 – bottom). Finally, we run SIDRA once more using the new light curve. We detect a planet with probability of 81 per cent.

We run the same experiment using BLS and it was able to detect the planet. We do not claim of course that this technique is new or it does not work with other detection algorithms. We just give an example showing that SIDRA is also able to detect exoplanets hidden in a strong variability.

4 CONCLUSIONS

SIDRA is a blind detection and classification algorithm based on Machine Learning – Random Forest technique. This paper is a general presentation of the algorithm. We used simulated light curves from five different classes. These are constant, transiting, variables, microlensing and eclipsing binary light curves. Assuming a 60 per cent decision probability, the algorithm success ratio is 95–100 per cent for microlensing, eclipsing binaries, variables and constant light curves and 91 per cent for transits for Table 1 input values. Also we test SIDRA with real light curves from the *Kepler* mission. We detect and classify successfully 100, 98 and 93 per cent of the constant, transiting and false alarm light curves. Furthermore, we show a simulated example of SIDRA transit detection around a variable host star.

We discuss the transiting exoplanet detection power compared with BLS-like algorithms, but we would like to make clear that we do not suggest to replace BLS with SIDRA, even if in our test SIDRA detects 8 per cent more planets and is 1000 times faster. What we suggest for a at least an exoplanet survey, is to include both algorithms. SIDRA could be a very powerful tool, and could easily detect/classify interesting objects which require further analysis.

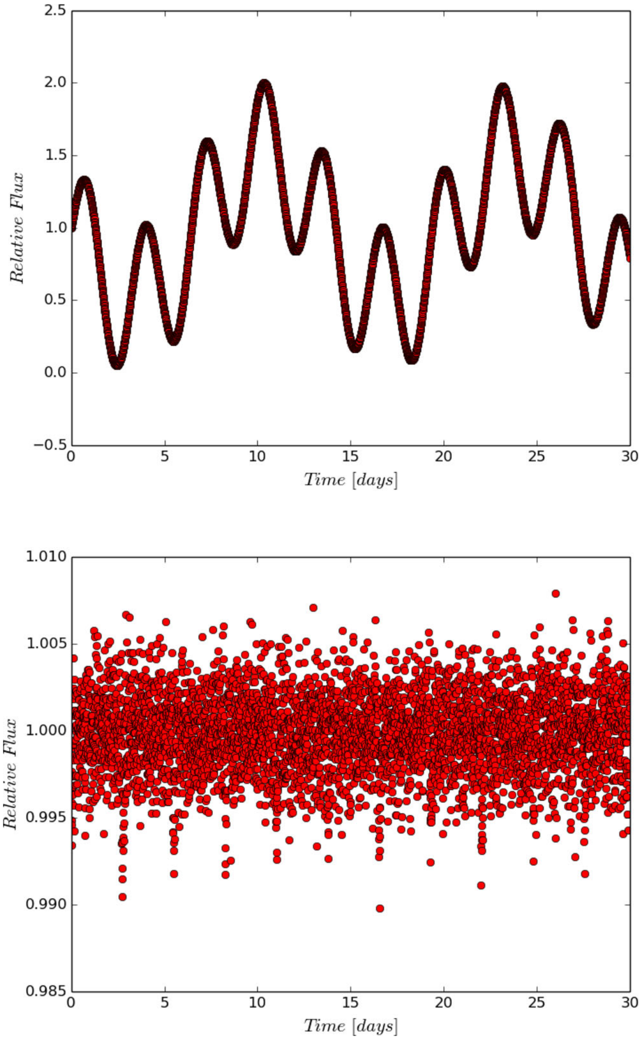

Figure 9. Top: a 30-d multiperiod variable star. Bottom: the same light curve after removing strong periodicities.

4.1 Advantages

The algorithm could be easily modified for each team and project and it is as general as possible, solving simultaneously different types of light curves. On the other hand, BLS or high $\Delta\chi^2$ methods for example, work only for transiting and eclipsing binary light curves.

For eclipsing binaries, variables and microlensing light curves, the algorithm is not only detecting the signal correctly but it also minimizes the false alarms. In our case of 10 000 simulated light curves, false alarms for these objects were eliminated.

For transiting light curves, it manages to detect more planets than the classical method of BLS. Furthermore, SIDRA is much faster than BLS. Once we train the network, the classification needs 4 ms for a single light curve of 4300 data points, making SIDRA ideal for huge surveys such as *TESS* and *PLATO* transiting exoplanet future space missions. BLS needs ∼4 s in the same machine for the same light curve.

Finally, SIDRA has the ability to become a huge network with almost all kinds of light curves. We can not only classify variable stars for example, but we can use many more classes in order to identify the type of each variable. On the other hand, we are able to search for non-periodic phenomena such as supernova and flare stars.

4.2 Disadvantages

The major disadvantage of the algorithm is that it cannot resolve physical characteristics from the light curves because of its nature. For example we cannot extract the information about the radius of the planet because we do not fit physical parameters but we calculate statistical values of each light curve. Of course for the periodic events, the information of the period is hidden in the autocorrelation function, but still there is more information which is missing.

Finally, the algorithm works with a training light-curve set, which means that we have to be very careful on the selection of features and limits in order to maximize the success of the algorithm.

4.3 False alarms

The main problem of every detection and classification algorithm is the false positive and negative alarms. Assuming a 20 000 light-curve sample, we expect from SIDRA to solve the majority of the light curves correctly. On the other hand, there are some limitations. We assume that the 1 per cent of the theoretical sample contains real planets, 15 per cent real variables, 10 per cent real eclipsing binaries and 0 per cent real microlensing events. Table 2 shows the results. From the total number the 95 per cent remain in our sample after the decision probability 0.6 cut. These remaining light curves (5 per cent) are flagged as unknown.

In order to deal with the 4 per cent of false alarm and the 5 per cent of unknown light curves, we can decrease step-by-step the decision probability from 0.6. This will increase the SIDRA sample, decreasing the unknown light curves. Also, we can use a typical BLS algorithm. It is not clear that BLS could solve the false alarms better than SIDRA (Fig. 6). That depends on the S/N of each light curve but we can use BLS as a separate tool. As we mention above, the decision probability is a very important variable and we plan to study it in a future publication.

ACKNOWLEDGEMENTS

This publication is supported by NPRP grant no. X-019-1-006 from the Qatar National Research Fund (a member of Qatar Foundation).

Table 2. Classes statistics assuming a 20 000 light-curve sample.

Classes	Total number	SIDRA sample (> 0.6)	Successful detection	False alarms	Unknown
Constant	14 800	14 060	13 357	703 as planets	740
Planet	200	190	175	15 as constant	10
Variable	3000	2850	2850		150
EB	2000	1900	1843	57 as planets	100
Total	20 000	19 000 (95 per cent)	18 225 (91 per cent)	775 (∼ 4 per cent)	1000 (5 per cent)

The statements made herein are solely the responsibility of the authors.

REFERENCES

Alcock C. et al., 2000, ApJ, 541, 734

Alsubai K. et al., 2013, Acta Astron., 63, 465

Bakos G., Hartman J., Torres G., Kovács G., Noyes R. W., Latham D. W., Sasselov D. D., Béky B., 2011, EPJ Web. Conf., 11, 01002

Batalha N. et al., 2013, ApJS, 204, 24

Bond I. et al., 2001, MNRAS, 327, 868

Bonomo A. S. et al., 2012, A&A, 547, 110

Bordé P., Fressin F., Ollivier M., Léger A., Rouan D., 2007, in Afonso C., Weldrake D., Henning Th., eds, ASP Conf. Ser. Vol. 366, Transiting Extrasolar Planets Workshop. Astron. Soc. Pac., San Francisco, p. 145

Borucki W. et al., 2010, Science, 327, 977

Breiman L., 2001, Mach. Learn., 45, 5

Cabrera J., Csizmadia Sz., Erikson A., Rauer H., Kirste S., 2012, A&A, 548, A44

Cameron C. A. et al., 2006, MNRAS, 373, 799

Claret A., 2004, A&A, 428, 1001

Foreman-Mackey D., Montet B., Hogg D., Morton T. D., Wang D., Schölkopf B., 2015, ApJ, 806, 13

Hogg D. W. et al., 2013, Kepler Project Office Call for White Papers: Soliciting Community Input for Alternate Science Investigations for the Kepler Spacecraft.

Kaler J. B., 1998, Mon. Notes Astron. Soc. South. Afr., 57, 89

Kovács G., Zucker S., Mazeh T., 2002, A&A, 391, 369

Li L., Zhang Y., Zhao Y., 2008, Sci. China G, 51, 916

McCauliff S., Jenkins J., Catanzarite J., 2014, A&AS, 766, 2

Masci F. J., Hoffman D. I., Grillmair C. J., Cutri R. M., 2014, AJ, 148, 21

Moutou C. et al., 2007, in Afonso C., Weldrake D., Henning Th., eds, ASP Conf. Ser. Vol. 366, Transiting Extrasolar Planets Workshop. Astron. Soc. Pac., San Francisco, p. 127

Mullally F. et al., 2015, ApJS, 217, 31

Paczynski B., 1986, ApJ, 304, 1

Pál A., 2008, MNRAS, 390, 281

Prša A., Zwitter T., 2005, ApJ, 628, 426

Robin A. C., Reylé C., Derrière S., Picaud S., 2004, A&A, 416, 157

Shannon C., Weaver W., 1949, The Mathematical Theory of Communication. Univ. Illinois Press, Urbana

MOA OGLE CollaborationSumi T. 2011, ESS.2.0103

Udalski A., Szymanski M., Kaluzny J., Kubiak M., Mateo M., 1992, Acta Astron., 42, 253

Wyrzykowski L. et al., 2014, Acta Astron., 64, 197

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