Earth Science Informatics
https://doi.org/10.1007/s12145-020-00464-7

REVIEW ARTICLE

Transiting Exoplanet Discovery Using Machine Learning Techniques: A Survey

Miguel Jara-Maldonado[1] · Vicente Alarcon-Aquino[1] · Roberto Rosas-Romero[1] · Oleg Starostenko[1] · Juan Manuel Ramirez-Cortes[2]

Received: 7 August 2019 / Accepted: 12 April 2020

Abstract

Spatial missions such as the *Kepler* mission, and the Transiting Exoplanet Survey Satellite (*TESS*) mission, have encouraged data scientists to analyze light curve datasets. The purpose of analyzing these data is to look for planet transits, with the aim of discovering and validating exoplanets, which are planets found outside our Solar System. Furthermore, transiting exoplanets can be better characterized when light curves and radial velocity curves are available. The manual examination of these datasets is a task that requires big quantities of time and effort, and therefore is prone to errors. As a result, the application of machine learning methods has become more common on exoplanet discovery and categorization research. This survey presents an analysis on different exoplanet transit discovery algorithms based on machine learning, some of which even found new exoplanets. The analysis of these algorithms is divided into four steps, namely light curve preprocessing, possible exoplanet signal detection, and identification of the detected signal to decide whether it belongs to an exoplanet or not. We propose a model to create synthetic datasets of light curves, and we compare the performance of several machine learning models used to identify transit exoplanets, with inputs preprocessed with and without using the Discrete Wavelet Transform (*DWT*). Our experimental results allow us to conclude that multiresolution analysis in the time-frequency domain can improve exoplanet signal identification, because of the characteristics of light curves and transiting exoplanet signals.

Keywords Artificial intelligence · Deep learning · Discrete wavelet transform · Exoplanets · Light curves · Machine learning · Multiresolution analysis · Transits

Introduction

The *Kepler* mission, launched on 2009, was aimed at monitoring the brightness of 150,000 stars in order to detect transits from Earth-sized and larger planets (Basri et al. 2005). This mission provided an excellent dataset of exoplanets, which are planets found outside the Solar System. Most of the data obtained by *Kepler* is publicly available through the Mikulsky Archive for Space Telescopes (*MAST*)[1]. Exoplanet research is important for several reasons:

- Obtaining statistics and information about the atmospheres and compositions of exoplanets and their host planetary systems.
- Extending our understanding on the mechanisms that built our own Solar System.
- Searching for habitable planets outside the Solar System (e.g. looking for planets at the habitable zone, i.e. at a distance in which liquid water could exist (Smith et al. 2012).
- Searching for life outside the Solar System. Nevertheless, no evidence of life has been found in the atmospheres of exoplanets, and it is not yet possible to take images of the surface of exoplanets (see (Seager and Bains 2015)).

Communicated by:H. Babaie

✉ Vicente Alarcon-Aquino
vicente.alarcon@udlap.mx

[1] Department of Computing, Electronics and Mechatronics, Universidad de las Americas Puebla, Sta. Catarina Martir, San Andres Cholula, Puebla Mexico 72810

[2] Department of Electronics, National Institute of Astrophysics, Optics and Electronics, Tonantzintla, Puebla Mexico 72840

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

Published online: 05 June 2020

Both scientists and volunteers (e.g. the Planet Hunters program, which has reported exoplanet discoveries such as the one in (Schwamb et al. 2013)) have been monitoring myriads of stars, with the objective of finding exoplanets in public datasets. This research initially required manual preprocessing of data to detect potential exoplanet candidates, and further determine if the analyzed candidate was a planet; or if the signal was caused by a different source. As pointed out by (Pearson et al. 2018), to examine all the data available manually is an extensive and exhausting task.

Some challenges in the area of exoplanet discovery, among others, are:

- Stellar variability, which causes variations in the brightness of the stars due to several factors (e.g. star spots).
- Noise sources that obstacle the discovery of exoplanets; such as pulsations, limb darkening (i.e. the effect that causes the center of the star to look brighter than its edge), outliers, systematic trends, and disturbing background signals, among others (see (Grziwa and Pätzold 2016) and (Smith et al. 2012)). Also, ground-based observatories have extra factors such as atmospheric seeing, bad weather, and lunar light pollution that difficult exoplanet discovery.
- Discontinuities within the light curves due to telemetry dropouts, spacecraft momentum dumping maneuvers, showers of solar protons during large solar flares, and others (see (Aigrain and Favata 2002)).
- False alarm rates related to manual misclassifications.
- The fact that transit signals from some planets are weak signals, sometimes too small and short to overcome noise.
- False positive sources such as Eclipsing Binaries (*EBs*, binary stars that orbit on our line of sight of its pair stars), stellar contamination, and in some cases even other planets found on our own Solar System (as it was the case of Mars within some observations from the *K2* mission ((Howell et al. 2014)), as explained in (Dattilo et al. 2019)), among others.
- The enormous amount of information available to discover exoplanet transit signals (e.g. the MAST Kepler Data Search page[2] contains nearly 3 million rows of links to different files related to the *Kepler* targets).
- The sensitivity requirements that the instruments must meet in order to detect exoplanets (e.g. in the transit method discussed later, the luminosity changes are often smaller than 1% as stated in (Yaqoob 2011)).

Over the past years, several Machine Learning (*ML*) and Artificial Intelligence *AI* approaches have been proposed to automatize the exoplanet discovery process. In some cases, these algorithms even allow one to detect shallow transits that are hidden by noise. ML can be used to reduce noise in the data.

In this work, a general overview on exoplanet research is provided. Also, we compare several works related to ML applied to transiting exoplanet research, with the objective of generating a comparative framework. Finally, we conclude that multiresolution analysis (*MRA*) techniques help to improve the exoplanet identification step. This survey is organized as follows. First, a general overview on exoplanets is given in Section "Exoplanet Overview". Then, in Section "Exoplanetary Detection Methods", the most important methods for exoplanet detection are explained. Section "Machine Learning Algorithms Used for Transiting Exoplanet Research" presents an analysis of ML algorithms used in transiting exoplanet research, organized according to the different steps of exoplanet discovery (discussed in that section). In Section "Multiresolution Analysis", MRA is proposed as a means to improve the detection and identification steps of the exoplanet discovery process, along with some other examples of its application to the general exoplanet research. We present our proposed model for synthetic dataset creation, along with our experimental results using MRA techniques and two synthetic datasets to improve the exoplanet identification performance in Section "Experimental Results". Next, Section "Discussion" includes a discussion about the reviewed ML approaches. Finally, conclusions are drawn in Section "Conclusions".

Exoplanet Overview

According to the International Astronomical Union (*IAU*), in its Resolution B5[3], the official definition for the word *planet* (which comes from the Ancient Greek for *wanderer*) refers to objects that comply with three properties:

- It orbits the Sun.
- It has enough mass to convey into a nearly round shape.
- It is the main gravitational object found in its orbit.

In the case of exoplanets, (short for extrasolar planet) they are planets found outside our Solar System, and the definition above does not entirely fit to them. As mentioned in (Khan et al. 2017), exoplanets are objects with masses lower than about 13 Jupiter masses (which is the limit before they start burning deuterium in their cores, thus becoming a brown dwarf, see (Burrows et al. 2001)), though large enough not to be called dwarf planets such as the case of Pluto. Also, there are exoplanets that do not orbit a star which are called isolated planetary mass objects (e.g. (Zapatero Osorio et al. 2000)). There is much discussion on

[2] https://archive.stsci.edu/kepler/data_search/search.php

[3] https://www.iau.org

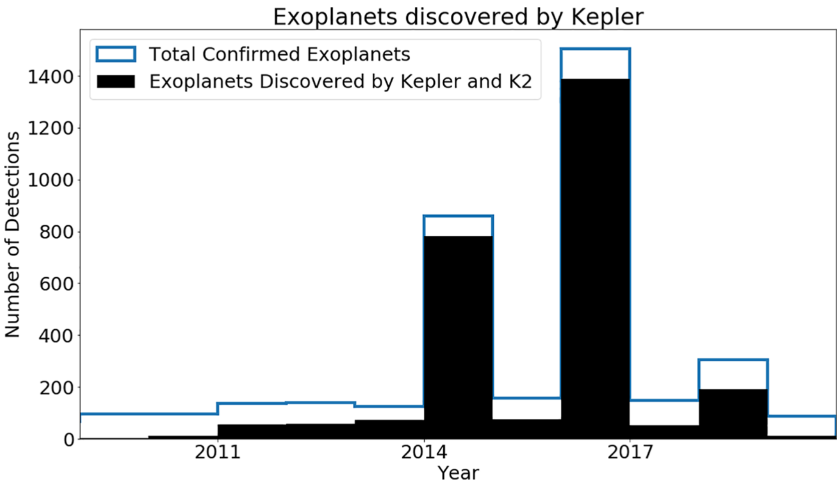

Fig. 1 Exoplanet discoveries up to May 2019. Line - Total number of confirmed exoplanets. Filled area - Number of exoplanets found with the *Kepler* and K2 missions. This figure was generated using data from the NASA Exoplanet Archive

which is the first exoplanet discovered. In 1988, (Campbell et al. 1988) reported an exoplanet called *Gamma Cephei A b* orbiting the Gamma Cephei star; but it was until 2002 when (Cochran et al. 2002) confirmed the detection of this exoplanet. Later, in 1989, a celestial body orbiting the star HD 114762 (considered as a probable brown dwarf) was reported in (Latham et al. 1989), and it has been added to the confirmed exoplanets list of the NASA exoplanet archive[4]. Nevertheless, some sources, such as (Yaqoob 2011), state that the first discovery was given in 1992, by Aleksander Wolszczan and Dale Frail, who discovered an exoplanet around a neutron star (Wolszczan and Frail 1992). Other sources, such as (Way et al. 2012), state that the first detection occurred in 1995 by Michel Mayor and Didier Queloz, by using the *radial velocity* method (Mayor and Queloz 1995) (this technique is discussed in Section “Exoplanetary Detection Methods”). Nonetheless, the exoplanet found by Michel Mayor and Didier Queloz, called “51 Pegasi b“, is the one that gave birth to the exoplanet research scientific field, because it is the first exoplanet discovered around a main-sequence star different from the Sun.

The first surveys to hunt transiting exoplanets were ground-based facilities such as the Wide Angle Search for Planets (*WASP*, (Pollacco et al. 2006)), and the Hungarian-made Automated Telescope (*HAT*)[5]. Then, space-based missions were also added to this cause, though several ground-based facilities are still being used for exoplanetary study. In 2009, the NASA launched the *Kepler* Space Telescope with the aim of discovering Earth-sized planets transiting Sun-like stars outside our Solar System in the habitable zone. It stared at a single field on the Cygnus constellation during four years, and the information that it retrieved is still being used to find more exoplanets, as it can be observed in Fig. 1. This figure presents a histogram of confirmed exoplanets published in the NASA Exoplanet Archive, found by using data from *Kepler*. According to (Jenkins et al. 2010), the *Kepler* data is divided into Long Cadence (*LC*) targets, which were sampled every 29.4 minutes; and Short Cadence (*SC*) targets sampled at intervals of 58.85 seconds. Once that the data was gathered by the *Kepler* satellite, it was received by the *Kepler* Mission Science Operations Center (*SOC*) through the Deep Space Network (*DSN*). The data were preprocessed by different modules of the SOC Pipeline, as asserted in (Stumpe et al. 2014), such as the Calibration module (*CAL*); the Photometric Analysis module (*PA*), which fitted and removed cosmic rays and sky background; and the Pre-search Data Conditioning (*PDC*) module, where systematic error sources, along with Sudden Pixel Sensitivity Dropouts (*SPSDs*) and other outliers were removed from the light curves. After data preprocessing, they were made publicly available through the Mikulsky Archive. As mentioned in (Chintarungruangchai and Jiang 2019), the publicly available data are divided into two different types of observed fluxes; namely the Simple Aperture Photometry (*SAP*) which contains the flux data with corrections for background flux, along with artifacts; and the PDC where artifacts have been removed. There are other modules used by the SOC to process the obtained data. The Transiting Planet Search (*TPS*) is used to detrend the light curves and to identify Threshold Crossing Events (*TCEs*), which are periodic decrements in the light flux that may be related to transiting planets. Then, the TCEs are evaluated by the Data Validation module (*DV*) to reject any false positives.

The *Kepler* mission recovered many Kepler Objects of Interest (*KOI*), which are stars that potentially host one or more exoplanets. There are several KOI catalogs that have been generated through ML algorithms and manual vetting (e.g. the catalogue of (Coughlin et al. 2016), among others). Furthermore, (von Essen et al. 2018),

[4] NASA Exoplanet Archive: http://exoplanetarchive.ipac.caltech.edu
[5] https://hatsurveys.org/

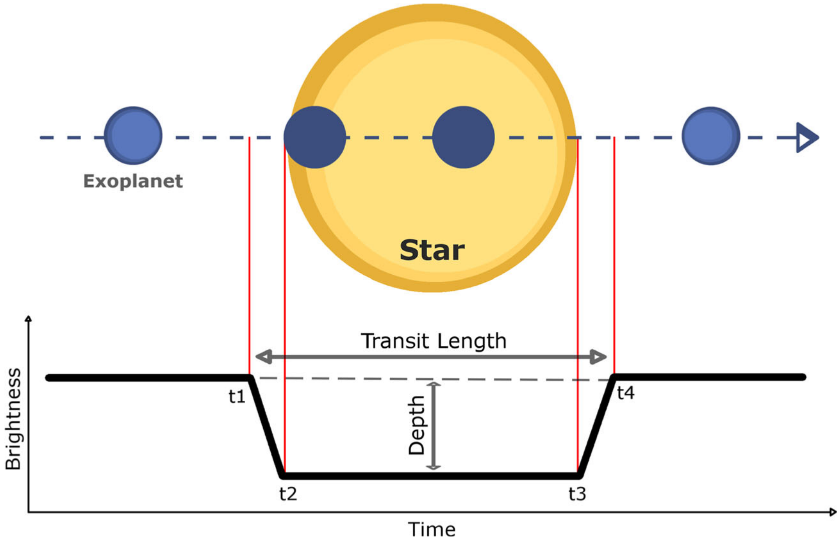

Fig. 2 Example of a light curve. As the exoplanet orbits the star, different brightness values are obtained. Some parameters that can be extracted from a light curve are: Beginning of ingress ($t1$); end of ingress ($t2$); beginning of egress ($t3$); end of egress ($t4$); transit length; and transit depth

and (Freudenthal et al. 2018), present the KOI Network (*KOINet*), which is a multi-site network composed by several telescopes spread over the world. The aim of the KOINet is completing the Transit Timing Variation curves (*TTV*, which is a deviation in the expected time of transit, caused by gravitational interaction in multi-planetary systems), where *Kepler* required additional data for a proper characterization. For more information on the *Kepler* mission, the reader is referred to (Jenkins et al. 2010), (Koch et al. 2010), and (Borucki et al. 2010).

When the *Kepler* spacecraft presented a failure in one of the reaction wheels that kept it pointed, the project turned into the *K2* mission ((Howell et al. 2014)). Using the remaining capabilities of the *Kepler* spacecraft, it could study young open clusters, bright stars, exoplanets, galaxies, supernovae, and asteroseismology. The definitive retirement of the *Kepler* observatory took place in the year 2018. Other telescopes for exoplanetary study are:

- Hubble Space Telescope: Even though this telescope was not designed to hunt exoplanets, according to (P Hatzes 2014), it was the first space-based facility to be employed in the study of exoplanets. Thanks to this telescope, it was formally accepted that space-based photometry could be enough to detect Earth-sized planets. It has been capable of studying the atmospheres of the exoplanets.
- CoRoT satellite (from the french *COnvection ROtation et Transits planétaires*, (Auvergne et al. 2009)): It is an European satellite which functioned from 2006 to 2012.
- Transiting Exoplanet Survey Satellite (*TESS*, (Ricker et al. 2015)): It was launched in 2018 with the goal of finding thousands of planet candidates transiting bright and nearby star. It uses the transiting photometry method (explained in Section 2).
- Other space-based telescopes such as the NASA Spitzer Space Telescope (*SST*, (Werner et al. 2004)), which was deactivated in 2020; and the CHaracterising ExOPlanets Satellite (*CHEOPS*, (Beck et al. 2017)), launched in 2019.
- Ground-based facilities such as the Wide Angle Search for Planets (*WASP*, (Pollacco et al. 2006)), Hungarian-made Automated Telescope (*HAT*), TRAnsiting Planets and PlanetesImals Small Telescope (*TRAPPIST*) project[6], and MEarth project[7], Atacama Large Millimeter/Submillimeter Array (*ALMA*, (Wootten and Thompson 2009)), among others.
- Future telescopes such as the Atmospheric Remote-sensing Infrared Exoplanet Large-survey (*ARIEL*, (Zingales et al. 2018)), James Webb Space Telescope (*JWST*, (Gardner et al. 2006)), Wide-Field Infrared Survey Telescope (*WFIRST*, (Pasquale et al. 2017)), PLAnetary Transits and Oscillations of stars (*PLATO*, (Rauer et al. 2014)), Thirty Meter Telescope (*TMT*, (Sanders 2013)), among others.

Exoplanetary Detection Methods

Astronomers have developed a wide variety of methods that can be used to look for exoplanets. These methods also allow one to determine different properties of the exoplanets (sometimes indirectly), such as the composition of the atmosphere of the exoplanet, its temperature, the mass of the host star and the exoplanet, among others. Even more, the combination of different methods may provide a better

[6] https://www.trappist.uliege.be/cms/c_3300885/en/trappist-portail

[7] https://www.cfa.harvard.edu/MEarth/Welcome.html

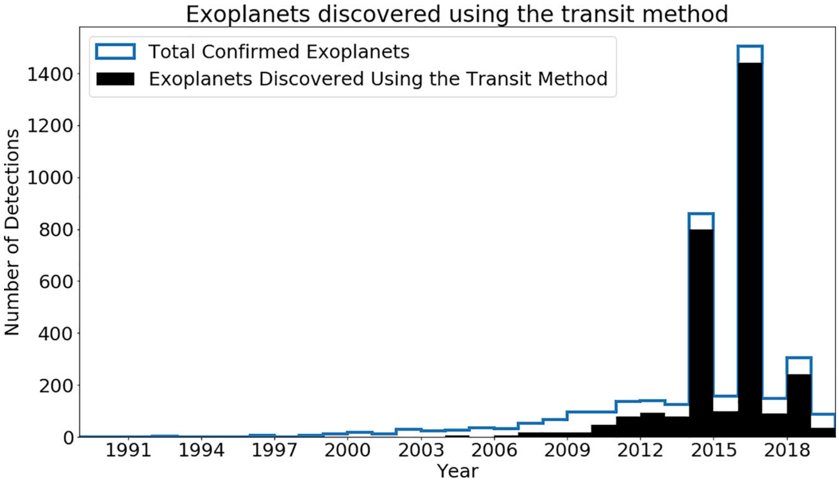

Fig. 3 Exoplanet discoveries up to May 2019. Line - Total number of confirmed exoplanets, regardless of the technique used. Filled area - Number of exoplanets found using the transit search method. This figure was generated using data from the NASA Exoplanet Archive

characterization of the exoplanet and star properties. Some of the most commonly used methods are discussed next.

Transit Method

A transit is a similar event to a solar eclipse, when an exoplanet passes between the observer and the star it orbits, it produces a transit. Transit events can be studied by using *light curves*; which are light intensity values as a function of time from the observed star. As the exoplanet passes in front of the star, the light curve shows a decrease in its brightness, meaning that there may have been a transit. The first success of this method was reported in 1999, with the discovery of the planet HD 209458b ((Charbonneau et al. 2000), previously discovered by (Henry and et al. 2000) using the radial velocity method). An example of an idealized light curve is shown in Fig. 2. The change in stellar flux during the transit event is very small (it is often $\sim$ 1% for hot Jupiters, and $\sim$ 0.01% for Earth-like planets, see (Grziwa and Pätzold 2016)). Furthermore, (Mandel and Agol 2002) have developed a model to simulate the stellar brightness during a transit. It has been used to create synthetic data for experimentation purposes in several works, such as (Pearson et al. 2018), (Shallue and Vanderburg 2018), (Foreman-Mackey et al. 2015) and (McCauliff et al. 2015) among others. The source code can be accessed through the web page of Eric Agol[8].

In some cases, the parameters that can be calculated through the transit method are the transit depth, transit length, ingress and egress times, ratio between the size of the planet and the size of its host star (R_p/R_*), orbital period (P), among others. Even more, there is an analytic solution that, under certain conditions, can be used to determine the stellar mass (M_*), stellar radius (R_*), planet radius (R_p), orbital semi-major axis (a), and orbital inclination (i) from a planet transit light curve (see (Seager and Mallén-Ornelas 2003)). According to the NASA Exoplanet Archive, up to 13 May 2019, the deepest exoplanet transit registered belongs to HATS-6 b ((Hartman et al. 2015)) which has a transit depth of 3.23 ± 0.03 (percent). Contrarily, the shallowest transit registered belongs to Kepler-37 b (Barclay et al. 2013), which has a transit depth of 0.0012 ± 0.0003 (percent). In the case of the transit duration, the exoplanet with the longest transit registered in the NASA Exoplanet Archive is Kepler-849 b ((Morton et al. 2016)), with a transit duration of 0.999654 ± 0.004154 (days). The shortest transit duration belongs to K2-266 b ((Rodriguez and et al. 2018)), which has a transit duration of $0.01389^{+0.00120}_{-0.00085}$ (days).

The main limitation of this method is that it can only be used to observe planets whose orbit plane is ”oriented so that it is observed edge-on, or nearly edge-on” ((Yaqoob 2011)), unless the planet is so close to its star that it shows a transit even at relatively high inclinations. This is statistically restricting for the method. Another limitation of this technique is that it is biased towards finding the largest exoplanets compared to the host star, as asserted by (Yaqoob 2011). There are other limitations of this method, for example, data gaps in ground-based measurements caused by the day-night cycle, among others. Finally, as shown in Fig. 3, most exoplanets have been found by using the transit technique. In addition, it is noticeable that the number of detections has spread out since the *Kepler* mission was launched in 2009.

A clearer example of a light curve is illustrated in Fig. 4; which corresponds to a real light curve from a planetary system around the star *HIP 41378*. It is a close up of the original light curve extracted from the K2 mission ((Howell et al. 2014)), and processed by the K2 Self-Flat-Fielding (*K2SFF*) pipeline (Vanderburg and Johnson 2014). For those readers interested in generating their own plots, it is recommended to follow the Transit Light Curve Tutorial[9].

[8] https://faculty.washington.edu/agol/transit.html

[9] https://www.cfa.harvard.edu/~avanderb/tutorial/tutorial.html

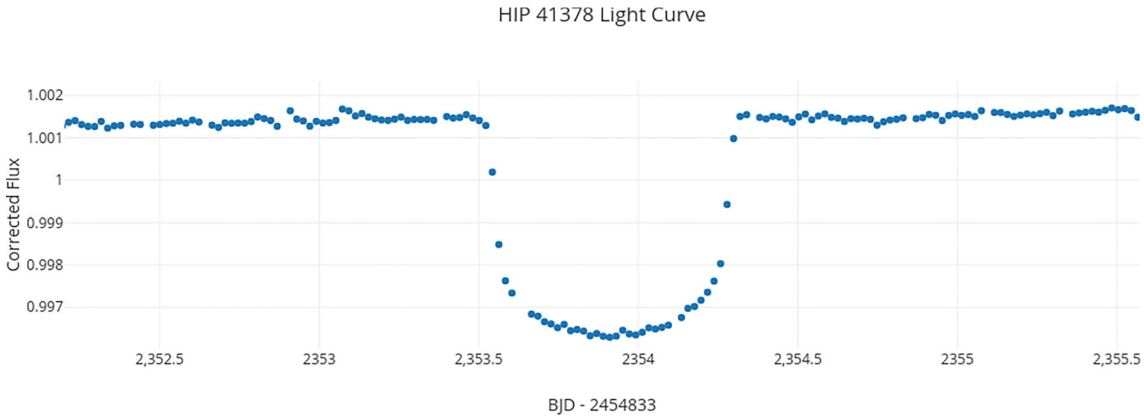

Fig. 4 Real light curve extracted from the planetary system around the star HIP 41378 in the MAST archive. The x-axis represents a measure of time called Barycentric Julian Day (*BJD*); the value 2454833 that accompanies the x-axis title, is to be summed to the x-axis value in order to calculate the BJD for each measurement. The y-axis represents the brightness of the star. This figure was created by following the Transit Light Curve Tutorial

Radial Velocity

This method consists in analyzing the Doppler shift effect observed in the host star, that is caused by the mutual gravity between the host star and the exoplanet (i.e. the stellar wobble). The stellar wobble can be represented by measuring the redshifts between two epochs. A redshift is when the observed light waves become more or less reddish as the star approximates and moves away from the observer. This approach is very used for exoplanet confirmation, and in many cases a combination of this method with the transit method provides a better characterization of the planet properties ((Yaqoob 2011)). Its main limitation is that stellar wobbles caused by exoplanets are too small which complicates their detection. Another limitation is that the spectrographs needed for this technique must be very stable ”at the centimeters/second level” for some Earth-sized exoplanets, see (Khan et al. 2017)), among other limitations. Readers interested in learning more about the radial velocity technique, as well as some signal processing approaches for this technique, are referred to (Khan et al. 2017), (Baluev 2018), and (Baluev 2013).

Gravitational Microlensing

Light trajectories are distorted by massive objects such as stars or planets. This distortion can change the direction of light, generating a gravitational lensing effect on the light of a star, as described in (Treu et al. 2012). The microlensing method relies on the fact that the gravity of an exoplanet can focus the light of distant stars to make them seem temporarily brighter. The main limitations of this method are that the required alignment of the star is unlikely to happen, and the fact that astronomers cannot predict where or when the lensing events will occur ((Yaqoob 2011)).

Direct Imaging

This technique consists in spatially resolving the exoplanet and its host star in order to obtain images from the exoplanets (under certain conditions). The image obtained is a small ”dot” as asserted by (Yaqoob 2011). Nevertheless, it can be used to obtain more details from the chemical composition and temperature of the exoplanet compared to the other methods. According to the NASA exoplanet archive, the first exoplanet found with this technique is called 2MASS J12073346-3932539 b ((Chauvin et al. 2004)), and it was discovered in 2004 orbiting a brown dwarf. The main issue with this technique, among others, is that it requires to construct instruments capable of spatially resolving the exoplanet from the star. One example of these instruments are the coronagraphs, such as Spectro-Polarimetric High-contrast Exoplanet REsearch (*SPHERE*, (Beuzit et al. 2019)), which have been designed to improve the image quality and contrast performance around bight stars. Also, exoplanets are very far away (and in some cases they are very small, e.g. Earth-sized planets). To exemplify the complexity of this technique, the closest exoplanet known up to October 2018 is *Proxima Centauri b*, which is found at 1.295 parsecs (*pc*) from Earth (i.e. 4.24 light years) as presented in (Anglada-Escudé et al. 2016). (Males et al. 2014) argue that the future generation of giant telescopes will be able to observe many exoplanets, but it is first required to achieve extreme contrasts at very small angular separations. The WFIRST telescope (which is expected to launch in the mid-2020s according to the project website[10]) is an example of a telescope that will enable exoplanet direct imaging.

[10] https://wfirst.gsfc.nasa.gov/

Exoplanet Discoveries

The aforementioned exoplanet detection techniques, as well as other techniques such as astrometry, have been used to generate a wide register of confirmed exoplanets. The NASA Exoplanet Archive is a dataset and tool-set funded by NASA and operated by the NASA Exoplanet Science Institute (*NExScI*). It includes over 2.9 million light curves from diverse projects such as *Kepler* and CoRoT. It also contains properties of the planets and their host stars, such as orbital periods, transit depths, and others. It is possible to obtain confirmed-planets and false-positive lists from these datasets using web-based tools, or by using *wget* and Hypertext Markup Language (*HTML*) calls. Furthermore, according to (Akeson et al. 2013), there are several criteria that an exoplanet observation must gather, before it can be submitted to the NASA Exoplanet Archive, namely:

- The exoplanet must have a mass estimate equal or less than 30 Jupiter masses.
- The properties of the planet must be described in peer-reviewed literature.
- Sufficient follow-up observations and validation must have been undertaken to deem the possibility of the object being a false positive.

From these points, it is possible to generalize the process of exoplanet discovery as:

- Data acquisition
  - Choose a detection method (e.g. the transit method). item[–] Build the instruments and missions to collect and store data related to the method selected (e.g. the *Kepler* mission recovers flux time series data that is used for the transit method). It is also possible to acquire the data by simulating it with models such as the ones reported in (Mandel and Agol 2002), (Kreidberg 2015), (Parviainen 2015) and (Emmanoulopoulos et al. 2013).
- Exoplanet discovery.
  - Look for a benchmark dataset to analyze (e.g. the MAST).
  - Preprocess the data (if it has not been already preprocessed) to remove the sources that are not related to exoplanet transits.
  - Apply a detection and later an identification model to the dataset (Such as the ones described in Section "Machine Learning Algorithms Used for Transiting Exoplanet Research").
- Discovery validation.
  - Confirm that the signal actually belongs to an exoplanet and not to a spurious detection (e.g. by performing an analysis with a second detection technique such as the radial velocity method as stated in (Moutou and Pont 2006)).
  - Submit the finding for peer review, to ensure its acceptance by the scientific community.

Table 1 Exoplanet detection techniques comparison

Technique	Percentage of detections
Transit.	77.8%
Radial Velocity.	18.2%
Microlensing.	1.9%
Imaging.	1.1%
Other methods.	1%

The NASA Exoplanet Archive gathers information by monitoring submissions via the journal pages and the Los Alamos National Laboratory (*LANL*) astro-ph server. Their dataset is updated on a weekly basis, that involves internal validation against the literature values.

Finally, Table 1 provides the exoplanet detection percentages for the techniques that have been discussed in this section (extracted from the Exoplanet Exploration NASA website[11]). As it can be observed, the highest number of detections has been obtained by using the transit method. In the following sections of this work, ML approaches that aim at discovering transiting exoplanets will be discussed.

Machine Learning Algorithms Used for Transiting Exoplanet Research

ML automatizes analysis tasks, so that computers may perform these tasks, which otherwise would need to be executed by humans. This enables scientists to analyze bigger quantities of information in less time. In this section, different ML approaches, that have been used for transiting exoplanet research, are explained and discussed. This section is organized according to a generalization of the exoplanet discovery process that we made based on the *Kepler* Pipeline work flow. The generalization is shown in Fig. 5; which describes 1) the light curve acquisition, 2) light curve preprocessing, 3) transit signal detection and 4) transit signal identification steps. The light curve acquisition step is not included in the analysis of this section, because ML focuses on processing the data that has already been obtained. The light curve preprocessing step is subdivided in two steps, the systematics and variability removal. However, the systematics removal step is only performed when the

[11] http://exoplanets.nasa.gov retrieved in 20/01/2019.

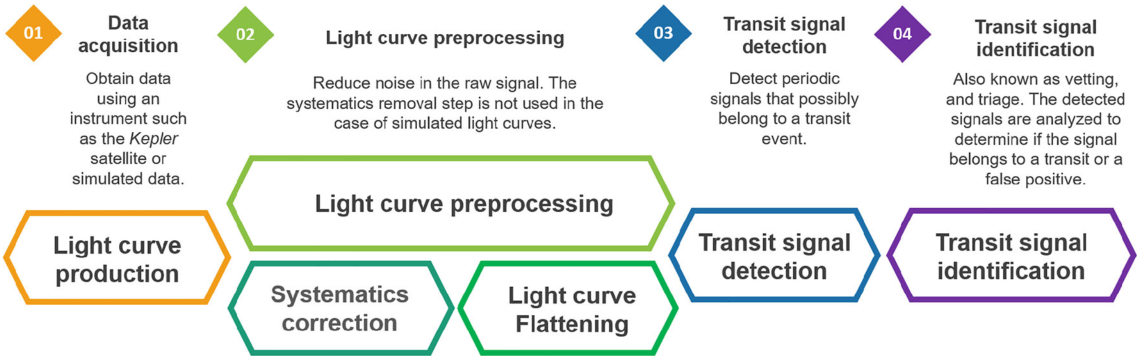
01
Data acquisition
Obtain data using an instrument such as the *Kepler* satellite or simulated data.
02
Light curve preprocessing
Reduce noise in the raw signal. The systematics removal step is not used in the case of simulated light curves.
03
Transit signal detection
Detect periodic signals that possibly belong to a transit event.
04
Transit signal identification
Also known as vetting, and triage. The detected signals are analyzed to determine if the signal belongs to a transit or a false positive.
Light curve production
Light curve preprocessing
Systematics correction
Light curve Flattening
Transit signal detection
Transit signal identification

Fig. 5 General flowchart of the transiting exoplanet discovery process

light curve is obtained from an instrument instead of being simulated. Also, notice that there is an important difference between the exoplanet detection and identification steps. In the detection step signals that could belong to an exoplanet transit are spotted, while in the identification step it is decided whether the detected signal belongs to an exoplanet or not.

Light Curve Preprocessing

The raw signals that are obtained by the observation telescopes (such as *Kepler*) contain several sources that could mask the transit signal (for example instrumental noise). Such sources may be reduced by applying certain preprocessing steps. As mentioned by (Aigrain et al. 2017), some light curves (such as SAP light curves) contain instrumental artefacts and systematic trends. Some of these trend sources are listed in (Jenkins et al. 2010); for example pointing errors, focus changes, thermal effects on instrument performance, and others. Additionally, some works (such as (Shallue and Vanderburg 2018)) also remove low frequency variability by applying a ”flattening” step. This step helps to eliminate natural variations in the light curve that are usually caused by stellar variability (perturbations on the flux that depend on the activity of each star), while preserving the shape of the transit (though this process may cause other issues such as the ones reported by (Shallue and Vanderburg 2018)). Even more, in order to give a greater importance to the shape of the transit (instead of the amplitude of the signal) some works, such as (Thompson et al. 2015), normalize the data so that no matter the event, the minimum and maximum flux values are always the same (e.g. minimum 0 and maximum 1). Finally, Table 2 lists several ML works. The models reported in such works are focused on the systematics correction and low variability removal of the light curves (i.e. the preprocessing step of exoplanet discovery). These works are discussed next. Also, the reader is referred to (Kovacs 2017), which contains more works regarding light curve preprocessing (including non ML models).

Least Squares

Grziwa et al. (2012) use a combination between the trend filtering algorithm from (Kovacs et al. 2005), and a Least Squares harmonic filter. The trend filter works by subtracting trends that are spotted on a subset of light curves from the same dataset to be used. The subtraction process has the advantage that there are no significant alterations to the transit signal while errors related to systematics are removed. In contrast, the Least Squares harmonic filter is used to remove periodic noise sources (e.g. stellar activity). This is done by fitting the harmonic sums of the dominant disturbing frequencies, using Least-Squares to remove them from the original signal. The transit signal remains unaffected after this process because it is not harmonic. Furthermore, this model is robust against data gaps; though it fails to remove noise sources with complex shapes. These filters are used to improve the detection rates of the EXOTRANS software package (discussed later), allowing one to detect signals with longer periods and dimmer transits.

Table 2 Machine learning models used for light curve preprocessing

Method	References
Least Squares.	(Grziwa et al. 2012)
Bayesian Approaches.	(Smith et al. 2012)
Dimensionality Reduction.	(Tamuz et al. 2005)
Multiresolution Analysis.	(Grziwa and Pätzold 2016), (Stumpe et al. 2014), (Carter and Nathan Winn 2009), (Jenkins 2002)

Bayesian Approaches

Smith et al. (2012) present a Bayesian approach based on the Bayesian Maximum a Posteriori (*MAP*), called Presearch Data Conditioning MAP (*PDC-MAP*). This model was used by the *Kepler* pipeline in the presearch data conditioning module. The aim of the model is removing systematic errors; while preserving sources of astrophysical interest such as transits or stellar variability. The motivation behind this model is that intrinsic stellar variability is not supposed to be correlated between different star data (because every star should have its own stellar variability); while instrumental noise should be correlated because all stars were observed by the same instrument. This assumption needs to be adjusted for highly variable stars and stars that present a *quite* behaviour. Thus, weighting constraints, granted by the MAP model, are used to trust the prior knowledge or not, based on the quantity of stellar variability presented by the target. On the one hand, if the target has a large stellar variability, the weighting constraints prevent the fitter from introducing noise and over-fitting the data. On the other hand, for low stellar variability targets, the prior knowledge is trusted.

To differentiate between highly variable targets and *quite* targets, and also to constrain the LS fit, the prior Probability Density Function (*PDF*) is parameterized according to three independent variables (viz. stellar magnitude, right ascension, and declination). By using these parameters, it is possible to find trends in the basis vector, which are generated by applying Singular Value Decomposition (*SVD*) to the most highly correlated targets. By using the prior PDF and the conditional PDF (both found fitting an LS model to the basis vectors), it is possible to obtain the posterior PDF that is measured based on the quality of the prior fit (called ”goodness” in this work), and the maximum likelihood function (which aims at finding the best LS fit with the least systematics). The main issues with this approach are that the SVD cotrending basis vectors often share noise from different sources, which difficults an independent removal of noise (and sometimes removes stellar variability). Also, some basis vectors inject high-frequency noise during the systematics correction (see (Stumpe et al. 2014)).

Dimensionality Reduction

The algorithm presented in (Tamuz et al. 2005) consists on an iterative process that removes systematics from the light curves, by searching a model that best fits the systematics. For this purpose, it uses Principal Component Analysis (*PCA*), which is a technique used for dimensionality reduction that finds the correlation between variables. The main advantage of this method is that it does not require a priori knowledge of the features related to the light curves. Nevertheless, it fails to detect all the systematics when they vary for each light curve; and in some cases, it may remove stellar variability.

Multiresolution Analysis

(Stumpe et al. 2014) present an improvement to the PDC-MAP model discussed earlier. It is based on a MRA methodology which they call multiscale MAP (*msMAP*). In this approach, the light curve is analyzed by obtaining different bands from the signals and correcting them accordingly with their resolution. Using different bands from a Discrete Wavelet Transform (*DWT*, using the Daubechies 12-tap wavelet as the mother wavelet) allows one to treat different kinds of systematics within the respective channels in which they are found. In this model, long-term trends are removed in band 1, artifacts of medium duration in band 2, and high frequency sources are treated in band 3. The MAP algorithm is used in the reconstructed version of each band (obtained with the reverse wavelet transform) in the time-domain. To enhance the representation of the noise within each band, channels are grouped together, giving robustness to the model in those cases where the noise source is present in more than one channel.

The msMAP approach implies introducing more parameters than the original PDC-MAP method; namely, the number of bands, the channel grouping boundaries (i.e. which channels to group together), and the possibility to set different PDF goodness parameters for each band (which enables the model to adjust to the necessities of each different band). As in the case of the PDC-MAP model, msMAP makes use of the goodness metric to evaluate the results obtained from the PDC cotrending process; though, in this case it is based in the removal of target-to-target correlations, injection of noise, preservation of stellar signals, and removal of Earth-point thermal recoveries. Using this information, the model can evaluate if msMAP has a better result than PDC-MAP, and in those rare cases in which msMAP fails ($1-2\%$ of the targets), PDC-MAP tends to give a better result and it is used. Nevertheless in a global scope, msMAP solves the residual systematic and high-frequency component injection issues from the PDC-MAP model, improving the quality of the output light curves.

The model for parameter estimation from a time series of exoplanetary transit photometry, proposed by (Carter and Nathan Winn 2009), is based on computing the likelihood in a wavelet basis. The purpose of this work is to estimate the mid-transit time t_c, by fitting a parametric model to a time series that may be contaminated by temporally correlated noise. The reason to estimate t_c is that it cannot be improved by using other transit events, and the variations in the transit interval could be due to other planets or satellites. Also, noise is characterized by using its Power Spectral Density (*PSD*). Furthermore, the Fast Wavelet Transform (*FWT*) is used to *whiten* the noise. The results prove that the proposed wavelet method works well in the presence of noise with short-range correlations. Furthermore, the authors state that one potential application of this method is the detection of transits in a photometric time series dataset.

(Grziwa and Pätzold 2016) use wavelets to create two filter methods, namely *VARLET* and *PHALET*. VARLET uses Stationary Wavelet Transform Denoising (*SWTD*) to decompose a signal by using wavelets. The decomposition allows one to extract stellar variability and discontinuities from high-resolution light curves. In concrete, SWTD consists in reducing noise in the time series by sequentially applying a series of filters (i.e., scaled wavelets) to a signal. In contrast, the PHALET filter is used to search for transits from multiple planets in the same light curve. It uses SWTD as well, but transits that have already been detected are extracted from the light curve. In addition, PHALET can be used to filter eclipsing binaries, which is useful to detect planets in binary systems.

The MRA model presented in (Jenkins 2002), uses an Overcomplete Wavelet Transform (*OWT*) to represent the light curve signal in both time and frequency. This model can treat nonwhite noise caused by stellar variability. Being able to analyse the signal in the time domain is important because stellar variability tends to change due to different factors such as increments in the magnetic activity of the star. The measurements used to represent stellar activity belong to measurements taken by the Differential Absolute RADiometer (*DIARAD*) instrument from the Solar and Heliospheric Observatory (*SOHO*) spacecraft ((Fleck 1995)). After applying the OWT to these data, a whitening filter is applied to each filter bank obtained by the wavelet, clearing out the noise found within windows of different sizes. Nevertheless, data gaps are a common problem in time series data. For this reason, the authors combine the data found within the gap edges; which could cause spurious information to be injected to the data, or even ignoring real transit signals.

Transit Signal Detection Models

The following works are related to the process of analysing the light curves and looking for signals that could belong to an exoplanet transit event. The output of these algorithms is usually a time reference to the moment in which the event was detected. The works analysed for such task are presented in Table 3.

Least Squares

(Kovács et al. 2002) use the Box-fitting Least Squares (*BLS*) algorithm to analyze stellar photometric observations in time series. Their aim is to look for periodic transits of exoplanets, by detecting *box-like* events in stellar light curves. The BLS algorithm aims at minimizing the Mean-Squared Error (*MSE*) between the expected data (predetermined box-like shape events) and the outputs. The main limitations of this method are that the signals have to be *box-like* in order to be detected; and it requires to have a balance between a good resolution of the box shape and the execution time (because the more resolution used to represent the signal, the more time the algorithm will need to execute). Also, an equation to characterize the Signal Detection Efficiency (*SDE*) is introduced in order to asses the BLS performance.

Grziwa et al. (2012) describe the *EXOTRANS* software package, which is used to detect transits after applying a harmonic and a trend filter. This work uses three types of BLS algorithms. The first is the BLS algorithm previously discussed (in (Kovács et al. 2002)), with the disadvantage that it discards transit signals found where there is an increase in the light curve intensity. For this reason, the second BLS implementation that they use is the directional BLS (*dcBLS*), which also reduces the false positive detections. Finally, the unmaximized BLS (*unmaxBLS*) presented in (Tingley 2003), is used to better fit the box shape into the transit shape. The problem with using the last BLS implementation is that the computation time considerably increases. For this reason, and because

Table 3 Machine learning models used in exoplanet detection

Method	References
Least Squares.	(Grziwa et al. 2012) (Kovács et al. 2002)
Dimensionality Reduction.	(Foreman-Mackey et al. 2015)
Bayesian Approaches.	(Carpano et al. 2003) (Aigrain and Favata 2002)
Match Filters.	(Petigura et al. 2013)
Deep Learning.	(Pearson et al. 2018), (Chintarungruangchai and Jiang 2019), (Zucker and Giryes 2018)
Multiresolution Analysis.	(Pearson et al. 2018)

the amount of data is huge (CoRoT produced 10,000 to 12,000 light curves sampled at 512s time interval with up to 25,000 data points each as mentioned in (Tingley 2003)), EXOTRANS works in a parallelized architecture of 20-50 processors , used to individually examine the light curves. Also, the search is limited to a range of periods between 0.5 and 15 days (varied 16,000 times). The authors explain that greater periods increase the possibility of detecting single events at wrong periods. In order to choose which periods to use, the authors used the SDE metric for 16,000 different periods. Finally, (Grziwa et al. 2012) assert that it is possible to detect weaker transit signals with longer orbital periods by using the trend and harmonic filters. This is because they are no more affected by the noise found within the light curves.

Dimensionality Reduction

In (Foreman-Mackey et al. 2015), a methodology where no systematics detrending takes place for detecting transit signals is proposed. The authors explain that detrending can cause over-fitting due to the reduction of the amplitude which can cause missing small signals coming from exoplanets. A PCA model is used to reduce the dimensionality of the stellar light curves, and exoplanets are searched as a linear combination of 150 parameters called Eigen Light Curves (*ELCs*). The methodology consists on three major steps. First, a likelihood function is applied to a set of different transit time and durations. Next, using the generated likelihood functions, the likelihood of a periodic model is now calculated based on the period, reference time, and duration of the transit. Finally, machine and human vetting is applied to discard spurious detections.

The choice of using 150 basis functions is arbitrary, and thus, optimality is not granted. Nevertheless, having a linear model allows one to perform optimization algorithms on a convex space. The likelihood function is conditioned to a specific set of periods, phases, and durations; which causes a possible large number of missing detections. In spite of that, this model can find transiting signals that were omitted by other transits. This is possible because it can remove the transit signals that have already been detected from the original signal. This enables the model to find more than one exoplanet transiting the same star. Finally, one of the main limitations of this approach is the false positive rate caused by EBs. Less than 10% of the total candidates are caused by EBs rather than exoplanet transits. Also, this model does not account for stellar variability.

Bayesian Approaches

Aigrain and Favata (2002) present a Bayesian approach, based on the Gregory-Loredo (*GL*) method, to detect planetary transits from terrestrial planets. The GL method was originally developed for the detection of pulsars (neutron stars that emit periodic radiation) in X-ray data. The new algorithm is built under the assumption that most planets will appear in the light curves of fainter stars. For this reason, it is necessary to construct a robust algorithm that can detect transits on noisy signals. The proposed modification of the GL algorithm can detect planetary transits by using bins of variable width. In this way, the phase of the transit can also be identified. The transit parameters used for such model are the period, duration and phase of the transit. This model can process signals with gaps of a few hours. Another advantage of this approach is that the model can approximate light curves with arbitrary shapes. Although the model should allow one to determine the transit duration, in practice it does not happen. (Aigrain and Favata 2002) attribute this to wide region fitting rather than the transit fitting.

Finally, as explained by (Aigrain and Favata 2002), the most serious noise that affects this method is likely to be intrinsic stellar micro-variability. For this reason, (Carpano et al. 2003) examine the impact of intrinsic stellar variability in (Aigrain and Favata 2002). They state that micro-variability is arguably the largest intrinsic noise source in transit searches, and that most critical noise sources are not white or well known. Even more, the authors demonstrated that removing stellar variability from the light curves is necessary for Bayesian algorithms. The proposal of (Carpano et al. 2003) is to use an optimal filter that aims to reduce the impact of Sun-like variability in simulated light curves. The *optimal filter* consists in the combination of an adaptive filter that whitens the noise, and a matched filter for the detection. Results demonstrate that without filtering, the detection performance decreases significantly.

Match filter

Petigura et al. (2013) measure the fraction of stars that host planets with sizes of 0.5 - 8.0 times the size of the Earth, and with orbital periods between 5 - 50 days. The search is restricted to a number of 12,000 stars (the *"Best12K"* sample) from the *Kepler* survey. The Best12k sample contains samples with the lowest photometric noise, with the aim of maximizing the detection of Earth-sized planets. Their model is called TERRA and it is an automated pipeline for small exoplanet detection that is composed of three steps. In the first step, median and high-pass filters are used along with a PCA model to calibrate the photometry in the time domain. In the next step, the Signal-to-noise ratio (*SNR*) is evaluated over a grid of different transit periods, epochs, and durations. If the maximum SNR peak is greater than 12 the signal is flagged for additional vetting. Finally, a data validation step is performed to fit the transit candidates with a transit model. During this last step, TERRA is used to vet *Kepler* TCEs. Even more, detection completeness is evaluated by injecting and recovering synthetic transits in the Best12K sample. The main limitation of this model is that it requires the transits to be strictly periodic, as well as knowing the period of the events in advance (which can be done by creating a grid of different periods). Also, the performance of the model decreases for low SNR transits, where variations in the time of the transit are longer than the transit duration. Furthermore, the data validation step still requires human intervention during the vetting process.

Deep Learning

A Convolutional Neural Network (*CNN*) used for exoplanet detection is presented by (Pearson et al. 2018). The authors state that deep nets learn to recognize planet properties instead of relying on hand-coded metrics that are prone to mistakes. The main advantage of using a deep net is the easiness to train the net with subtle features in large datasets. The deep net is trained with simulated noisy photometric data created by using a grid of different parameters; such as the transit depth and orbital period. Finally, the deep net uses both transit and non-transit cases to learn how to distinguish one from another. Even more, this work presents a comparison among different classifiers including their CNN, a wavelet Multilayer Perceptron (*MLP*), a Support Vector Machine (*SVM*) and the BLS algorithm (among others). Their results show that the 1 Dimensional Convolutional Neural Network (*1-D CNN*) obtains the best performance values. They conclude that their proposal could be ameliorated with the use of feature transformations such as wavelets, which would be used to discard the least significant features in the light curves. Similarly, (Zucker and Giryes 2018) present a model based on a 1D-CNN used to detect transit signals in noise induced simulated light curves that have not been detrended. As mentioned by (Zucker and Giryes 2018) , simulated data can prove useful to train the ML models, because they guarantee which light curves contain transits and which ones do not.

A proposal to use a two dimensional CNN (*2D-CNN*) that takes into account a new way of folding the light curves for exoplanet detection is presented by (Chintarungruangchai and Jiang 2019). The authors of this work use two transit periods instead of one for the folding process, causing the resulting light curve to have all periods folded within two points. Their proposed model is compared against five different methods, including an MLP, two 1D-CNNs and two 2D-CNNs to analyze the effects of using different versions of the input light curves, namely using the whole light curve as input, folding the input light curves using the traditional method, and using the proposed two transit folding method. Another innovation of this work is that instead of using a single vector containing the average of all the folds of the light curve, they use a 2D-CNN. With the 2D-CNN architecture they can create a grid with each folding step on each row as an input (similar to having a pixel grid). This allows their model to search for transits when the signal is not exactly folded according to the transit period, which is usually the case when the transit period is unknown. This helps to reduce the resolution of periods to be tested because the transit periods will be represented in the 2D matrix (i.e. the grid); whereas the average of the folds could cancel the transit period by averaging it with the other transits that are out of phase. Furthermore, in order to test the reliability of the 2D-CNN against the other five models, different tests are performed by varying the SNR, transit phase positions, and the folding points.

The results obtained by (Chintarungruangchai and Jiang 2019) show that this model is robust against noise because the folding process enhances the transit signal (it obtained an accuracy, precision and recall above 98% with a *log*(SNR) lower than 1.0). Also, the model is robust to different transit phase positions because of the new folding process. Finally, (Chintarungruangchai and Jiang 2019) tested their model by varying the folding points to consider situations where the light curves are not folded according to the transit period. The maximum variation between the folding period and the transit period in their experiment was of 20%. When the model was trained with the period differences, the model obtained an accuracy of about 95%. The 2D-CNN is thus robust to low SNR and to those cases where the transit period is unknown. Nevertheless, the signals of transits belonging to small planets are ignored in this work, which would be an interesting amelioration to the current model, since it has the capability of detecting transit signals where the SNR is low.

Multiresolution Analysis

(Pearson et al. 2018) train a neural network by using the detail and approximation coefficients of the DWT of light curves as inputs. By doing this, the algorithm can learn the most significant pieces of the signal, while ignoring the least important ones, such as noise. For this purpose, a Fully Connected (*FC*) deep net was designed, which used input data treated by a wavelet transform (viz. the second order Daubechies wavelet (Daubechies 1992)). In concrete, the detail coefficients (*cDs*) and approximation coefficients (*cAs*) were appended together, to use them as inputs for a wavelet MLP. Their results demonstrate that the wavelet MLP has a training accuracy of 99.77%, and a sensitivity of 91.50%; which demonstrates the efficiency of the algorithm. This could be due to the fact that wavelets adapt to different shapes. As the authors state, exoplanet transits have different shapes and a simple template cannot be enough to recognize subtle details, specially in the cases where the signal is hidden by noise or strong systematics are present. One limitation of this approach is that, by appending both cAs and cDs, subtle features that are only present on one type of coefficients could be masked by the features from the other type of coefficients. Their algorithm, along with a Support Vector Machine (*SVM*), 1-D CNN, and others, can be retrieved form their github repository[12].

Transit Signal Identification Models

Once that a transit-like event has been detected, it is necessary to decide what is the exact phenomenon that has been detected. The following works are focused on determining if an event belongs to a transiting exoplanet, or to another kind of source. This process is also known as vetting because it involves doing a revision of the detected events, and vetting those that do not belong to an exoplanet signal. Normally, the output of these algorithms is a yes or no answer to the question ”Is the detected event caused by an exoplanet?”. Nevertheless, some identification algorithms can also determine the source of false positive signals. The works related to this process are presented in Table 4, and are discussed next.

Decision Trees

(Coughlin et al. 2016) present an exoplanet identification model called *Robovetter*. The proposed approach consists in performing several tests, by using decision trees. Decision trees work by partitioning the space of attributes several times, trying to reduce the entropy within each partition step until the entropy is no more reduced. The terminal partitions are called *leaf nodes*, and once they are obtained, the algorithm can predict the class associated to each leaf node. Robovetter decides if a candidate signal falls into one or more of the following four False Positive (*FP*) categories:

- Not Transit-like: TCEs where the light curve does not correspond to a transiting planet or EB.
- Significant Secondary: A TCE that has a significant secondary event, which indicates that the transit-like event could be caused by an EB.
- Centroid Offset: A signal from a TCE whose origin could be a nearby star, rather than the target star.
- Ephemerides Match: A TCE with the same period and epoch (ephemerides) as another object.

Robovetter makes decisions based on the period of the TCE, Multiple Event Statistic (*MES*) and maximum Single Event Statistic (*SES*) values, and the radius of the planet (which is calculated by multiplying the radius ratio, obtained from the data validation module of the *Kepler* pipeline, by the stellar radius value). Finally, the catalog generated by their model is compared to several other catalogues, obtaining an overall agreement rate greater than 95%.

(Catanzarite 2015) proposes a model called *Autovetter*. This model implicitly maps values from a set of attributes to one of three different classes, by using a decision tree approach called random forest. The Autovetter has two refinement steps; the first consists in bootstrapping the training dataset (i.e. creating several trees to build a forest, by varying the samples on the training set), this refinement step enables to avoid over-fitting the algorithm (by training the algorithm on a broader variety of environments); and the second refinement is achieved by choosing different small random subsets of attributes to decorrelate the trees that will be created. The three classes used by Autovetter are Planet Candidate (*PC*), which corresponds to signals consistent with transiting planets; Astrophysical False Positive (*AFP*), for those signals of astrophysical origin that could mimic planetary transits; and Non Transiting Phenomenon (*NTP*), which are spurious signals attributed to noise sources.

The inputs of the algorithm are the training set which consists of TCEs classified in the previous three classes by human disposition; and the attributes that can be extracted from a list of 114 possible attributes (e.g. period, transit depth, transit epoch, stellar parameters, etc.). The importance of each attribute is measured to provide the most significative ones on a list that can be consulted in (Catanzarite 2015). This work also provides a Bayesian estimate of the posterior probability that a TCE is member of each class, which is used to calculate the confidence that the model had on the classification process. Finally, (Catanzarite 2015) compares the results of the Autovetter with the Robovetter. They conclude that even though both approaches differ on certain decisions, for the most, they agree on 87.9% of the times that Robovetter classifies a signal as a PC and 97.2% of the times that the Autovetter classifies a signal as a PC.

(Armstrong et al. 2018), present a ML algorithm that uses a random forest, and Self-Organizing-Maps (*SOMs*). This work is focused on the vetting process of signals obtained by the Next Generation Transit Survey (*NGTS*), which is a ground-based survey presented in (Wheatley et al. 2018). Random forests use a set of Decision Trees to perform classification; in this case, a set of input features that were extracted from the light curves are used (e.g. the transit shape statistic, transit duration, etc.). Even more, it is possible to reduce the variance and bias of the model by using a diversity of trees. Each different tree has its own subset of random features. The final output probability is the fraction of trees that decided on each class, which can be used to perform candidate ranking. Also, feature importance is obtained by observing the size of the tree: The

[12] https://github.com/pearsonkyle/Exoplanet-Artificial-Intelligence

Table 4 Machine learning models used for exoplanet identification

series Method	References
Decision Trees.	(Catanzarite 2015), (Coughlin et al. 2016), (Armstrong et al. 2018), (Schanche et al. 2019)
Dimensionality Reduction.	(Thompson et al. 2015)
Deep Learning.	(Dattilo et al. 2019), (Shallue and Vanderburg 2018), (Ansdell et al. 2018), (Yu et al. 2019)

higher the tree, the more influence a feature has on the final classification.

Additionally, the SOM algorithm is an unsupervised learning approach (i.e. that it does not require a labeled training set), that clusters groups of inputs based on Euclidean distance between the input points. The SOM designed by (Armstrong et al. 2018) consists in a 20 × 20 grid of ”pixels” (called *Kohonen layer*); which is used as a randomly initialized transit template. During the training step of the SOM algorithm, transit templates are permuted to obtain different shapes in the inputs (by using different periods, epochs, and transit durations). The training process involves finding the best matching pixel in the layer, by minimizing Euclidean distance between pixel elements and the input. The SOM algorithm can separate planetary candidates from a wide range of noise sources, as well as astrophysical and instrumental phenomena. The algorithm implementation of (Armstrong et al. 2018) can be downloaded from their github repository[13]. Also, synthetic injected planets are used in order to guarantee survey independence. Finally, the authors demonstrate that the features concerning the transit depth are important for eclipsing binary detection; but also affect the identification of several known planets with deep transits. For this reason, they have used two models: One that considers the transit depth (With Depth-Related Features *WD-RF*), and one that does not (Non Depth-Related Features *ND-RF*).

Schanche et al. (2019) perform a comparison among several models used to analyze stellar light curves from the WASP archive. From the different models tested (viz. SVM, Logistic Regression, linear SVM, *k*-nearest neighbors, and the Random Forest Classifier), the random forest obtained the best performance. Given a set of 4,627 samples for training, and 2,280 for testing; the random forest model obtained a recall of 94, which means that it could find a wider range of planets, albeit its false positive number was high, having 137 FPs and 45 correct planet identifications. For this reason, the authors decided to train a Convolutional Neural Network (*CNN*). Results lead to conclude that while the CNN had a better accuracy, their Random Forest model could recover several planets missed by the CNN.

Some other works that have used Random Forest for light curves classification can be found in (McCauliff et al. 2015), (Armstrong et al. 2015) and (Nun et al. 2014).

Dimensionality Reduction

(Thompson et al. 2015) introduce a new metric for transiting exoplanet signals identification. This metric uses the Locality Preserving Projections (*LPP*) dimensionality reduction method, along with *k*-nearest neighbors to look for transit-like shapes within the light curves. Using LPP enables the algorithm to reduce the dimensionality of the data, be more robust to outliers than other methods (such as PCA), and to cluster transit-like signals by measuring the Euclidean distance between each data point and its *k*-nearest neighbors. In concrete, the metric proposed is the mean of the *k* distances between each neighbor. A TCE is considered to belong to a transit if the LPP metric shows that the transit was found close to other transit-like events, in a space where transit-like and non-transit-like events are separated from each other (which is the dimensionality reduced space of 20 dimensions generated by using the LPP algorithm presented in (He and Niyogi 2004)).

The data used for this approach corresponds to the TCE list reported in the Q1-Q17 *Kepler* Data Release 24 *DR24* ((Coughlin et al. 2016)). These light curves are detrended using two different method (viz. the Data Validation median filter (*DV-median filter*), and the penalized Least Squares (*penalized-LS*) method). Then, the authors fold the light curves, centering the TCE at a phase 0.5. Next, the folded light curves are binned according to the transit event; in other words, the data is grouped into 141 bins, 51 bins corresponding to in-transit points and 90 to out-of-transit to encourage the algorithm to use more or less the same number of data points related to the transit. Having bins outside the transit allows the model to account for large variability and low harmonics of the detected period; whereas the in-transit points enhances the transit signal. Finally, in order to base the model in the shape of the transit (instead of its amplitude) the data is normalized.

This model uses TCEs with a MES $>$ 8 to create the sample transit and non-transit like signals; which means that they discard small planets found within signals with

[13] https://github.com/DJArmstrong/TransitSOM

low SNRs for their model. Also, the usage of the *k*-nearest neighbors clustering algorithm means that if there are transits that vary in shape from those used to calculate the Euclidean distances, they will not be identified by this model. Nevertheless, the authors analyzed the results from using different detrending methods and they concluded that alternating between both of them enables the model to detect transit-like signals within different situations. For example, the penalized-LS can track transits within variable stars that are discarded by the DV-median method. The MATLAB implementation of the LPP metric used in (Thompson et al. 2015) is available on SourceForge[14].

Deep Learning

In (Shallue and Vanderburg 2018) a CNN called *AstroNet*, as well as other models, are used to classify potential planet signals. The tested architectures are an Artificial Neural Network (*ANN*) with zero hidden layers, a Fully Connected Neural Network, and a 1-D CNN that assumes that input light curves can be described by spatially local features, and that outputs do not change if there are small translations of the input. The output of the three models is the predicted probability that the input data belongs to a transiting planet. When the output values are close to 1 it means that the model has a high confidence that the signal corresponds to a transiting exoplanet. Instead, when values are close to 0, the signal is more likely to be a false positive. Furthermore, the proposed models consider three different input options, namely the local view (which focuses on the transit signal), global view (which includes data away from the transit) and a combination of both local and global view; which are used to represent short and long period TCEs. Results show that the best model (from the three aforementioned models) is the CNN. For such reason, the CNN is used to automatically vet *Kepler* TCEs, by using TCE labels from the Mikulsky archive for Space Telescopes of the *Kepler* mission with three possible values: Planet Candidate (*PC*), Astrophysical False Positive (*AFP*), and Non-Transiting Phenomena (*NTP*); which are then turned into planet and non planet labels. Even more, this work compares the CNN model against other vetting algorithms over simulated data, which are summarized in Table 5. The other algorithms are:

- Robovetter ((Coughlin 2017)).
- The Autovetter in (Catanzarite 2015): In this case, the results were not comparable to the CNN model because of the architecture of the model itself.
- Unsupervised ML in (Armstrong et al. 2016): It consists in a ML algorithm that clusters light curves with similar shapes.

Table 5 Vetters accuracy comparison against the CNN presented in (Shallue and Vanderburg 2018)

Algorithm	Accuracy	Accuracy of the CNN
Robovetter.	0.974	0.960
Autovetter.	0.986	–
Unsupervised.	0.863	0.949

This model is very confident with signals with MES values greater than 10 (where MES are constructed from single event statistics time series, that represent the likelihood of a transit of the fitted duration as a function of time; and are used to determine the statistical significance of the transit sequence, as explained in (Jenkins et al. 2015) and (Coughlin et al. 2016)). However its identification performance decreases when the signals have a MES value between 5 - 10 (It is likely that Earth-sized planets will be found near within these values, but it is also the *Kepler* pipeline detection limit). One of the main limitations of this model is that, during the signal preprocessing step, light curves are flattened and, occasionally, this creates false transits in stars with high-frequency stellar variability. Nevertheless, two new planets are statistically validated, and the validation process is also presented on their work. Furthermore, the CNN can identify exoplanets on multi-planet systems (e.g. the two planets validated by their work). Even more, the CNN can accurately distinguish subtle differences between the transiting exoplanets and false positives caused by instrumental artifacts, EBs, and stellar variability. The source code of their CNN can be found at their github repository[15]. Similarly, (Dattilo et al. 2019) use the base architecture of the AstroNet for the identification process in data from the *K2* mission, called *AstroNet-K2*. This work presents several modifications to the original AstroNet, such as its adaption to a different dataset (the *K2* mission data), and the use of different information related to the transit event as scalar inputs for the CNN (such as the planet/star radius, and the transit impact parameter). Similarly to AstroNet-K2, the *AstroNet-Triage* model proposed by (Yu et al. 2019) is a modification to the AstroNet applied to data obtained from the TESS mission.

Finally, (Ansdell et al. 2018) made an improvement to the model of (Shallue and Vanderburg 2018) (the CNN previously discussed). The modification consisted in adding scientific domain knowledge to the model architecture. In concrete, the centroid time series information obtained from the *Kepler* data, as well as some of the key stellar parameters, are used as inputs for the CNN. This allows the CNN to identify false positives such as background

[14] https://sourceforge.net/p/lpptransitlikemetric/code/HEAD/tree/

[15] https://github.com/google-research/exoplanet-ml

eclipsing binaries. The stellar parameters used are the stellar effective temperature (T_{eff}), surface gravity ($log\ g$), metallicity, radius (R_*), mass (M_*) and density (ρ_*); which are extracted from the updated *Kepler* DR25 catalogue in (Mathur et al. 2017). Also, they alleviated the model overfitting by implementing data augmentation techniques. The improved model is called ExoNet, and a simplified version called Exonet-XS is also presented, which has a lower number of convolutional layers for the local and global views, while maintaining the improved performance. The improvements increased the model accuracy and average precision by $\approx\ 2.0 - 2.5\%$. It is remarkable that they obtained $15 - 20\%$ gains in recall to the lowest signal-to-noise transits, which may correspond to rocky planets in the habitable zone. The source code is available in gitlab[16].

Multiresolution Analysis

When trying to discover new exoplanets, the noise found within the light curves imposes a detection limit and raises the probability of having false positives. An autonomous algorithm for exoplanet detection or identification, should be capable of extracting subtleties from the light curves by itself. The aim of this section is to recommend the use of MRA for exoplanet research, especially during the exoplanet detection and identification steps. For this reason, the wavelet theory is briefly explained; and some examples of wavelets used in exoplanetary science are presented.

Wavelets

According to (Veitch 2005), wavelets are wave functions that grow and decay over a finite time interval. Wavelets are used to localize a given function in both position and scaling. They are usually denoted by $\psi(\cdot)$. The $\psi(\cdot)$ function (called mother wavelet) is used to create several wavelets by translating and dilating it, as Eq. 1 shows.

$$\psi_{\lambda,\tau}(u) = \frac{1}{\sqrt{\lambda}}\psi\left(\frac{u-\tau}{\lambda}\right) \tag{1}$$

where $\lambda > 0$ is the dilation parameter; and τ is the translation parameter. Wavelets are convolved with the signal (with each modified scale controlled by the dilation parameter) to determine how much does a section of the signal resemble the wavelet. Some examples of wavelet models that have been used in general in exoplanet research (i.e. not necessarily for a certain task or exoplanet detection method) are presented next.

Examples of Wavelet Models Used in Exoplanetary Science

Masciadri and Raga (2004) propose a wavelet analysis technique for automatic exoplanet recognition, using direct imaging observations of nearby stars from ground-based and adaptive optics facilities. The aim of this work is to recognize exoplanets from a deep image obtained from the Very Large Telescope (*VLT*) infrared camera (*CONICA*, (Lenzen et al. 1998)), and the Nasmyth Adaptive Optics System (*NAOS*, (Rousset et al. 2000)). To achieve this task, they use a particular set of spectral coefficients to build *template planets*. These templates are used to look for features with the same spectral coefficients in the deep image. The template planet is then added to the deep image of the central star. Then, the central star image and the added template planets are convolved with a set of wavelets with different scales. The model marks all the points that have the same spectrum defined by a tolerance value. If a point is marked but does not coincide with the template planets, it is considered as a possible planet-like object. However, in the case of planet-like features, further work has to be done to discriminate if a planet has been found, or if they belong to a background star. Similarly, (Bonse et al. 2018) use wavelets for speckle-suppression with the aim of improving the SNR achievable for ground-based exoplanet imaging assisted by adaptive-optics.

Bravo et al. (2014) use the sixth-order Morlet wavelet to identify the temporal evolution of different phenomena that affected light curves, such as transiting exoplanets. A time-frequency analysis was performed, by using the wavelet transform on different light curves from the CoRoT and *Kepler* missions. The aim of this work is to identify particular properties associated to rotation, magnetic activity and pulsation. By applying a convolution between the wavelet and the signal, it was possible to determine the resemblance of a section of the signal with the selected wavelet. Results demonstrate that wavelets offer a detailed interpretation of light curves, providing more information of the different physical phenomena found within the signal.

Experimental Results

We have created two data sets in order to test different ML models. The aim of these experiments is to determine the performance of each model; and also to evaluate if there is an improvement in their performance by using MRA. Some of these ML models have been used in the works presented

[16] https://gitlab.com/frontierdevelopmentlab/exoplanets

Table 6 Transit simulation parameters

Parameter Name	Range of values
Orbital period (P).	0.0253 – 46.69 days
Orbit semi-major axis (a/R_*).	0.0058 – 0.2535 AU
Orbit inclination (i).	78.3 – 96.5 deg
Planet radius (R_p).	0.063 – 1.98 Jupiter radii
Orbit eccentricity (e).	0 – 0.53
Stellar radius (R_*).	0.12 – 2.59 Solar radii
Transit depth (R_p^2/R_*^2).	0.0085 – 3.23%
Transit duration.	0.0253 – 0.4113 days
Argument of periastron (ω).	90
Mid transit time (t_0).	75 days
Transit resolution.	150 data points
Limb darkening model.	Uniform, linear, quadratic and nonlinear
Limb darkening coefficients (u_1, u_2, u_3, u_4).	[empty], [0.5], [0.5, 0.1], [0.5, 0.1, 0.1, -0.1]

in Section 2, though we use our own implementations without the enhancements presented in each work.

Data Handling

Our two datasets consist on 10,000 light curves of 15,000 data points each. The main difference between both datasets is that for the first one (from now on referred to as the *Real-LC dataset*) we used 10,000 real light curves (obtained from the MAST) that were marked as NTPs in the Q1-Q17 *Kepler* DR24 ((Coughlin et al. 2016)). Instead, the second dataset (from now on referred to as the *3-median dataset*) was built by using simulated noisy photometric data. In both cases, we injected simulated transits to the half of the light curves using the BATMAN model from (Kreidberg 2015). The parameter grid used for the transit simulations was selected according to a list of 140 real exoplanets retrieved from the NASA Exoplanet Archive (see Table 6). We chose the only 140 exoplanets in the DR24 that were discovered using the transit method, and that had certain parameter values needed to simulate the transits with the BATMAN model (or at least their estimates). These parameters are the orbital period (P), orbit semi-major axis (a/R_*), orbit inclination (i), planet radius (R_p), orbit eccentricity (e), stellar radius (R_*), transit depth (R_*), and transit duration. In some cases, it was necessary to perform the necessary convertions to the units reported in the archive, so they could be fitted into the BATMAN model (e.g. the planet radius, which is reported in Jupiter radii units and must be converted into Stellar radii units). Also, we used four variations of the limb darkening model for each of the 140 exoplanets to obtain a wider variety of exoplanets. The total number of simulated transits that we created is thus 560. These 560 simulated transits were injected to 5,000 light curves both in the Real-LC dataset and the 3-median dataset.

To create the Real-LC dataset (see Fig. 6) we obtained the 10,000 real light curves. Then, we detrended the light curves by using the spline fitting method from (Shallue and Vanderburg 2018). This process consists in fitting a spline to the light curve, and then dividing the light curve by the best-fit spline to remove low-frequency variability. The result is a ”flattened” version of the light curve. Next, we iteratively added the 560 simulated transits until 5,000 light curves contained transit signals. The transit resolution

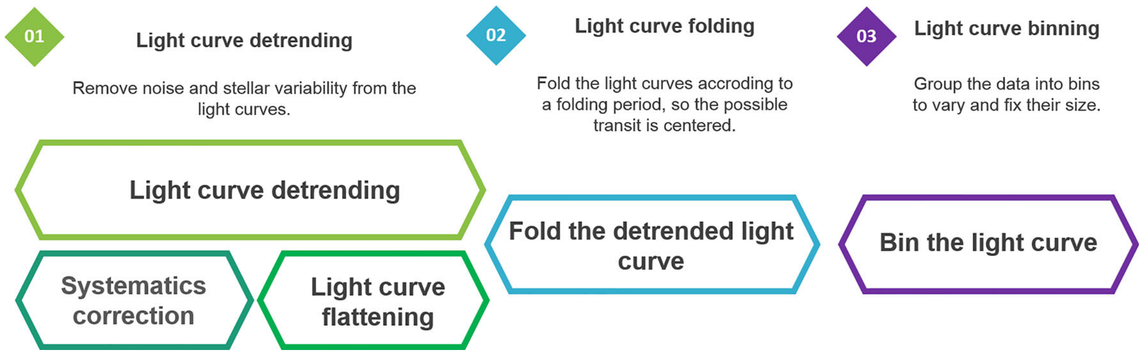
01
Light curve detrending
Remove noise and stellar variability from the light curves.
02
Light curve folding
Fold the light curves accroding to a folding period, so the possible transit is centered.
03
Light curve binning
Group the data into bins to vary and fix their size.
Light curve detrending
Systematics correction
Light curve flattening
Fold the detrended light curve
Bin the light curve

Fig. 6 General machine learning input preprocessing pipeline

was set to 150; therefore, we had to repeat the transit signal 100 times in order to fill the whole 15,000 data points light curve. We applied linear interpolation to the light curves that presented any data gaps. Then, we proceeded with the folding and binning steps. The folding step helps to enhance the transit signal. This step traditionally consists in folding the light curves according to a folding period. Then, all the folds are added by calculating the average of all folds. The resulting light curve contains one transit that represents all the transits of the original light curve. We used the PyAstronomy AstroLib FoldAt function for the folding process, which also normalizes our flux data so that the values range from 0 to 1. Finally, the binning step is used to vary and fix the size of the data that is going to be used as inputs for the ML models. This step consists in grouping the data into ”bins” that contain, for example, the mean of all the data points within one bin so that its size is reduced. We opted for a length of 2048 bins, which is similar to the 2001 bin global view used by (Shallue and Vanderburg 2018). This length was chosen so that we could apply DWT with a maximum of six levels of decomposition to the light curves (for each level of decomposition the signal looses half its size; thus, at six decomposition levels we have 32 data points left).

The 3-median dataset was created differently (see Fig. 6). First, we simulated 5,000 transits using the same method as in the Real-LC dataset. Then, we used Eq. 2 (Pearson et al. (2018)) to simulate 10,000 noisy light curves, from which 5,000 contained the simulated transits.

$$t' = t - t_{min}$$
$$A(t') = A + A\sin\left(\frac{2\pi t'}{P_A}\right)$$
$$\omega(t') = \omega + \omega\sin\left(\frac{2\pi t'}{P_\omega}\right)$$
$$F_{tr}(t) * \mathcal{N}\left(\frac{R_p^2}{R_*^2}/\sigma_{tol}\right) * \left(1 + A(t')\sin\left(\frac{2\pi t'}{\omega(t')} + \phi\right)\right) \quad (2)$$

where $F_{tr}(t)$ is the transit simulated using the BATMAN model from (Kreidberg 2015), t is the time, t_{min} is the first time value, σ_{tol} is the noise parameter, A is the amplitude of the simulated stellar variability, ω is the period of oscillation, ϕ is the phase shift, $\mathcal{N}$ is a Gaussian distribution with a mean of 1 and standard deviation of $(R_p^2/R_*^2)/\sigma_{tol}$, P_A and P_ω allow one to configure the frequency and amplitude of the simulated variability, and R_p^2/R_*^2 is the normalized radius ratio between the planet and the star. Non-transit noisy data was generated using the same Equation, but without the transit signal $F_{tr}(t)$, as reported by (Pearson et al. 2018). The parameter grid that we used to create the 10,000 light curves is presented in Table 7. After generating the noisy light curves, we detrended them by using a 3 median filter; but instead of dividing the original signal by the 3 median filter, we used the result of the filter as the new light curves. Finally, we applied the folding and binning steps to these light curves.

Table 7 Noisy light curve simulation parameters

Parameter Name	Range of values
Noise parameter (σ_{tol}).	0.25, 0.75, 1.25, 1.75, 2.25, 2.75, 3, 10
Wave amplitude (A).	0.025, 0.05, 0.1, 0.2
Wave period (ω).	6./24, 12./24, 24./24
Phase offset (ϕ).	0
Amplitude variability period (P_A).	−1, 1, 100
Wave variability period (P_ω).	−3, 1, 100

We tested the performance of several models using different data as inputs. In concrete, we used the binned light curves without any modification, and the DWT of the binned light curves as inputs for the ML models tested. In the case of the DWT, we used the cDs alone or the cAs alone in six different decomposition levels as inputs (from the first to the sixth level of decomposition). According to (Alarcon-Aquino and Barria 2009), the wavelet should be selected based on how well it adapts to the event to be analyzed, in this case the transit signals. In order to have a better representation of the transit signal, we selected several wavelets with different vanishing moments. Moreover, we chose orthogonal and bi-orthogonal wavelets for comparison purposes. The wavelets that we used were the daubechies 1 (*db1*), daubechies 5 (*db5*), symlet 5 (*sym5*), coiflet 5 (*coif5*) and bi-orthogonal 2.4 (*bior2.4*) wavelets. Furthermore, the number of scales (i.e. the number of decomposition levels) depends on the overall energy displayed at each scale ((Alarcon-Aquino and Barria 2001)). We chose six different decomposition levels to evaluate if different transit signals could be present at different scales. The light curve DWT process is shown in Fig. 7, which

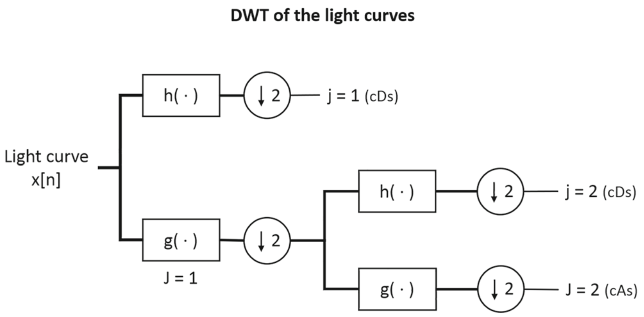

Fig. 7 Wavelet or detail coefficients (cDs) and scaling or approximation coefficients (cAs) are obtained by applying a series of high-pass filters ($h(\cdot)$) and low-pass filters ($g(\cdot)$) to the original light curve signal $x[n]$. Also, j denotes the cDs, while J denotes the cAs; and they indicate the number of decomposition levels

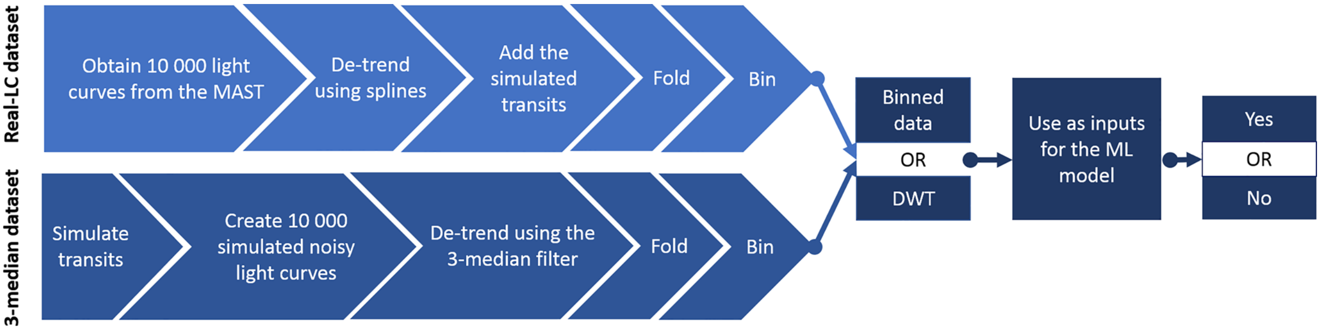
Real-LC dataset
Obtain 10 000 light curves from the MAST
De-trend using splines
Add the simulated transits
Fold
Bin
3-median dataset
Simulate transits
Create 10 000 simulated noisy light curves
De-trend using the 3-median filter
Fold
Bin
Binned data
OR
DWT
Use as inputs for the ML model
Yes
OR
No

Fig. 8 Dataset creation proposed model. The steps from above correspond to the Real-LC dataset, while the steps from below correspond to the 3-median dataset. In the final steps, the Discrete Wavelet Transform (*DWT*) coefficients of the binned data, or the data without DWT are used as inputs for the Machine Learning (*ML*) model

indicates that the original light curve signal is decomposed in cAs and cDs at each level of decomposition.

Finally, the creation process of both datasets is summarized in Fig. 8. As it can be observed in this figure, the last three steps correspond to the ML model input, execution, and output process. Finally, an example of applying DWT (using the db1 wavelet) to a binned light curve from the Real-LC dataset is shown in Fig. 9. Also, as a result of the downsampling process the number of data points is reduced by half on each decomposition level.

Model Setting

Using the datasets presented in the previous section, we have trained 11 different ML models to identify which light curves have simulated transits and which do not. The models that we have trained are six different configurations of a MLP, a SVM, a CNN, a Random Forest, a Naive Bayes, and a Least Squares (*LS*) model. We used the scikit-learn python API to program our models, excepting the CNN which was implemented by using the TensorFlow python library. We opted for six different MLPs in order to test the impact in performance of having different hyper-parameter configurations. All the MLPs had a learning rate of 0.001, $\alpha = 0.0001$ (i.e. the L2 penalty parameter), a maximum of 200 iterations, a tolerance of 0.0001, the number of iterations without change was 10 (the training stops before the 200 iterations if the score is not improved by at least the tolerance value for 10 epochs), and a lbfgs solver function. The architectures of the first two MLPs consist of two hidden layers with five and two neural units; and one architecture uses the sigmoid function as activation (Sigmoid MLP(5, 2)), while the other uses the relu function for activation (Relu MLP(5, 2)). The next two (Sigmoid MLP(1024), and Relu MLP(1024)) only had one hidden layer of 1024 neural units. We did this accordingly

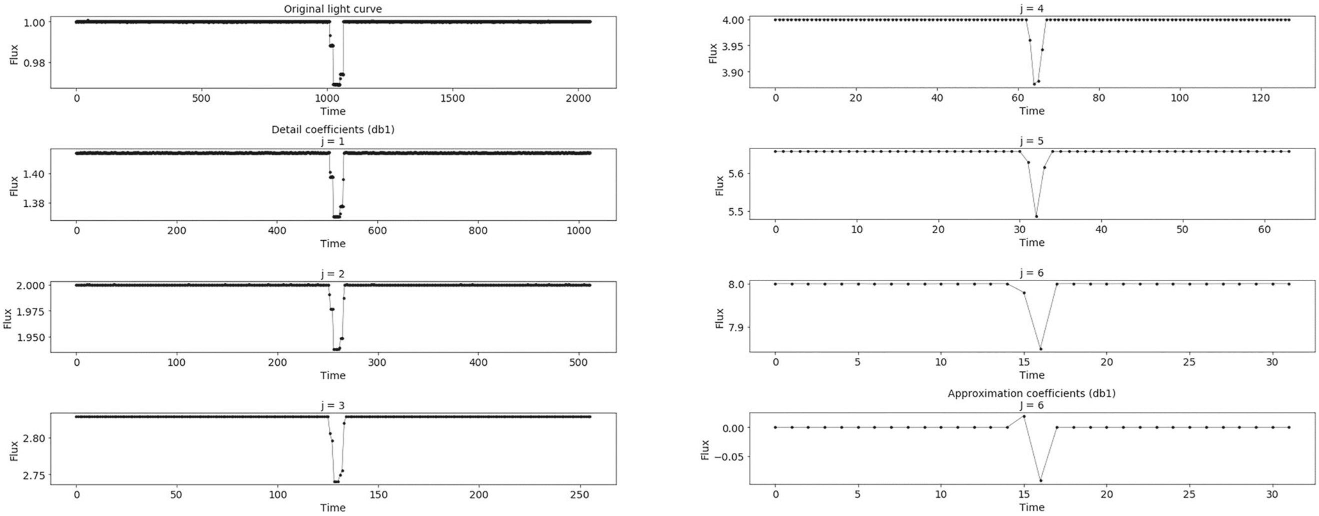

Fig. 9 Example of the detail and approximation coefficients that are obtained after applying the DWT to a binned light curve using the db1 wavelet. The detail coefficients are denoted by j, while the approximation coefficients are denoted by J; and they indicate the number of decomposition levels

with the description of (Alarcon-Aquino and Barria 2006), which says that the number of hidden units should be ”approximately equal to half the sum of the number of input and output units” (i.e. the half of 2048 inputs and one output). The last two MLPs (Sigmoid MLP(64, 32, 8, 1), and Relu MLP(64, 32, 8, 1)) consisted in four hidden layers with 64, 32, 8, and 1 neural units accordingly. These last two MLPs are based on the description of (Pearson et al. 2018), although the model reported in that work used a different learning rate.

The Random Forest had 10 trees in the forest with no limit of expansion (i.e. until the branches were completely expanded into leaves), the split quality criterion was set to the Gini impurity criteria, the minimum number of samples to split a node was 2, the minimum samples on a leaf were 1, and the trees were built by bootstrapping. We used the Gaussian Naive Bayes with no specified prior probabilities of the classes. The LS model was configured using a linear regression classifier, without fitting the intercept.

The SVM was configured with a regularization parameter of 1.0, a radial base function (*rbf*) kernel, a scale kernel coefficient, a tolerance of 0.001, and no limit of iterations. We based our CNN in the model proposed by (Shallue and Vanderburg 2018), using only the global view. All the hidden layers use the Rectified Linear Units or linear rectifier (*ReLU*) activation function, except for the output layer which uses a sigmoid activation function. The only limitation of using this architecture is that it can only be used with two decomposition levels at the most. Recall that the inputs length is reduced by half on each decomposition level due to the downsampling process of the DWT; thus, at more than two decomposition levels the inputs length is too short to undergo the downsampling caused by the five max pooling layers. For this reason, we modified the architecture of the model for the last 3 - 6 decomposition levels.

All the CNN architectures are shown in Fig. 10, to which we set the batch size of the CNN to 64 with 50 epochs. We used the Adam optimization algorithm ((Kingma and Ba 2014)) with $\alpha = 10^{-5}$ (i.e. the stepsize), $\beta_1 = 0.9$, $\beta_2 = 0.999$ (which are the exponential decay rates), $\epsilon = 10^{-8}$ (which is used to avoid dividing by zero during the parameters update), and a categorical cross-entropy loss function.

Simulation Results

We have trained and tested each model (excepting the CNN) 100 times with every possible input configuration (i.e. using

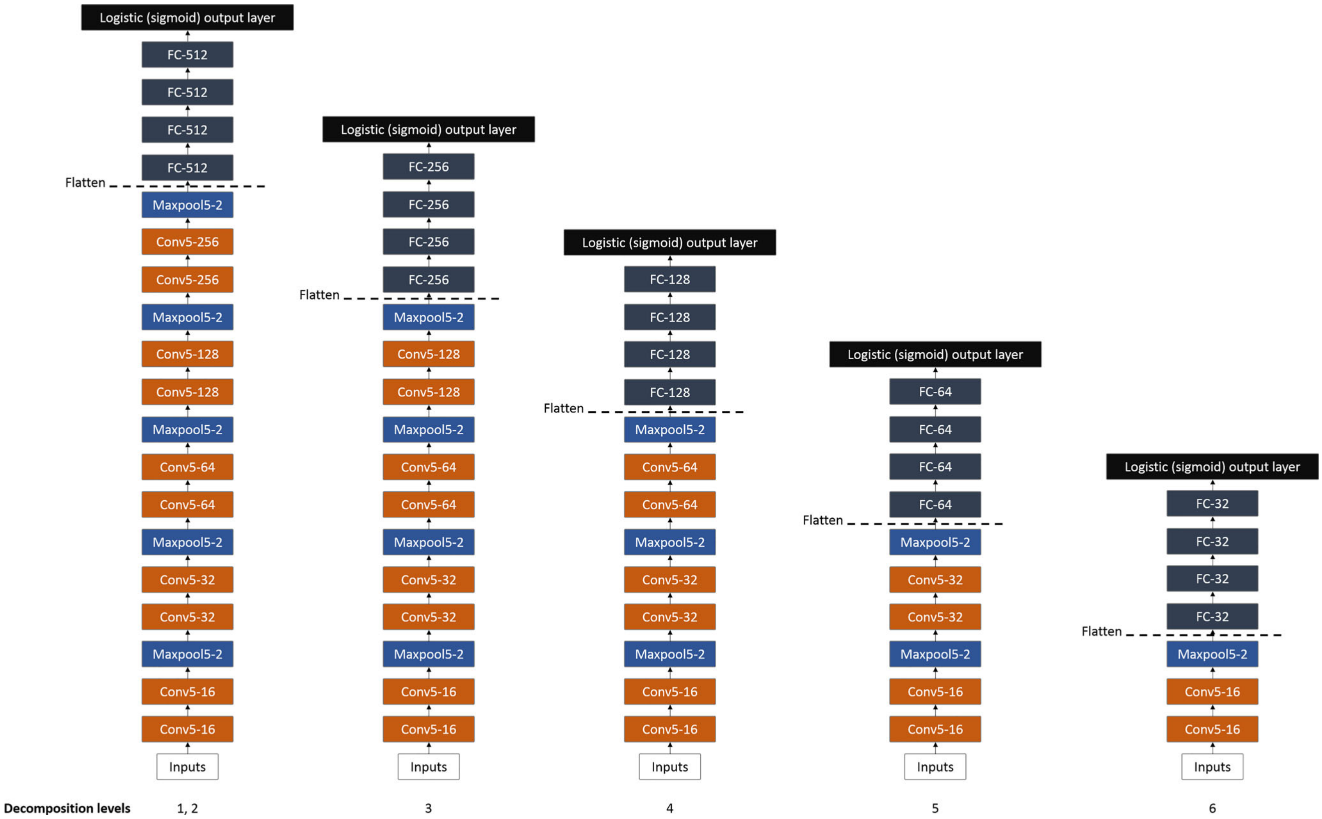

Fig. 10 Architecture of the Convolutional Neural Network (*CNN*) according to each decomposition level of the Discrete Wavelet Transform (*DWT*). Convolutional layers are described as Conv[kernel size]-[filters], max pooling layers as Maxpool[pool size]-[strides], and Fully Connected (*FC*) layers as FC-[number of units]. The number of decomposition levels is found beneath its related architecture

the binned inputs without MRA, and using the DWT cAs or cDs on six decomposition levels for every wavelet). In the case of the CNN, we only trained the model once and then tested it 100 times. This was done with the purpose of avoiding adding excessive training times to the results. We have used the scikit-learn python API *train_test_split* function on each of the 100 executions with all the models (including the CNN). This function allowed us to divide and shuffle our dataset into 60% data points for training, and 40% for testing the models. For each different input configuration, we measured the average accuracy, precision, recall, specificity and execution time of the 100 executions. The aforementioned metrics can be used to evaluate the performance of a ML model, based on the number of *yes* answers that were correctly classified (True Positives, *TPs*) and incorrectly classifed (False Positives *FPs*), and the number of *no* answers that were correctly classified (True Negatives, *TNs*), and incorrectly classified (False Negatives, *FNs*). The accuracy measures the percentage of times that the model was correct, the precision is used to measure how often a *yes* answer given by the model was correct, the recall or true positive rate measures the percentage of real

Table 8 Real-LC dataset experimental results, averaged from 100 executions of the training and testing processes for each ML model. The ML model inputs are based on using the DWT coefficients with several wavelets and decomposition levels, and without applying DWT to the binned light curves

ML model	Accuracy (%)	Precision (%)	Specificity (%)	Recall (%)	Time (seconds)
Sigmoid MLP(5, 2)					
Inputs without DWT.	49.65	28.83	52.0	48.39	4.82
Inputs using db5 cDs ($j = 6$).	92.0	94.76	94.92	89.08	0.67
Relu MLP(5, 2)					
Inputs without DWT.	49.54	21.26	57.0	43.0	5.75
Inputs using db5 cDs ($j = 6$).	83.87	98.84	99.16	68.55	0.31
Sigmoid MLP(1024)					
Inputs without DWT.	88.73	98.85	98.48	79.0	49.57
Inputs using coif5 cAs ($J = 2$).	88.79	99.18	98.33	78.27	24.14
Relu MLP(1024)					
Inputs without DWT.	81.63	93.19	87.02	76.23	84.44
Inputs using sym5 cDs ($j = 6$).	89.12	97.63	97.93	80.36	30.06
Sigmoid MLP(64, 32, 8, 1)					
Inputs without DWT.	49.49	30.2	41.0	59.0	5.05
Inputs using db5 cDs ($j = 2$).	49.5	27.7	44.0	56.0	1.41
Relu MLP(64, 32, 8, 1)					
Inputs without DWT.	76.98	86.71	88.56	65.48	15.42
Inputs using db5 cAs ($J = 2$).	77.88	93.54	87.81	67.87	6.81
LS					
Inputs without DWT.	65.16	94.26	98.04	32.23	8.51
Inputs using bior2.4 cAs ($J = 2$).	62.32	98.08	99.50	25.21	1.8
Random Forests					
Inputs without DWT.	97.91	98.35	98.37	97.45	10.26
Inputs using db1 cAs ($J = 4$).	98.50	98.59	98.59	98.41	1.18
CNN					
Inputs without DWT.	91.46	97.55	91.46	85.21	46.74
Inputs using coif5 cAs ($J = 2$).	94.28	98.15	94.28	90.16	31.59
Naive Bayes					
Inputs without DWT.	53.78	70.28	86.11	21.76	6.17
Inputs using db1 cDs ($j = 6$).	77.09	94.16	92.53	61.64	0.16
SVM					
Inputs without DWT.	88.67	99.34	99.49	77.81	60.87
Inputs using db1 cDs ($j = 6$).	93.06	98.56	98.72	87.4	0.78

yes answers that were classified as such, and the specificity or true negative rate measures the percentage of real *no* answers that were classified as such (see Equations 3, 4, 5 and 6 from (Japkowicz and Shah 2011)).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

We report the results without using DWT and the best results obtained by DWT in Tables 8 and 9. The best results for each model (between using the DWT of the binned data and the binned data without DWT as inputs) are highlighted in bold letters. Also, we use j and J to denote the coefficients that give the best results. The j notation refers to the cDs; whereas the J notation refers to the cAs. Additionally, we included the decomposition level that obtained the best results next to the corresponding

Table 9 3-median dataset experimental results, averaged from 100 executions of the training and testing processes for each ML model. The ML model inputs are based on using the DWT coefficients with several wavelets and decomposition levels, and without applying DWT to the binned light curves

ML model	Accuracy (%)	Precision (%)	Specificity (%)	Recall (%)	Time (seconds)
Sigmoid MLP(5, 2)					
Inputs without DWT.	84.64	77.43	76.31	92.79	12.42
Inputs using coif5 cDs ($j = 4$).	94.23	92.38	91.69	96.72	1.76
Relu MLP(5, 2)					
Inputs without DWT.	49.98	42.67	15.0	85.0	7.65
Inputs using bior2.4 cDs ($j = 3$).	95.12	92.94	92.27	97.93	2.54
Sigmoid MLP(1024)					
Inputs without DWT.	86.04	79.85	72.87	98.93	166.42
Inputs using bior2.4 cAs ($J = 6$).	92.65	89.72	88.13	97.1	46.94
Relu MLP(1024)					
Inputs without DWT.	93.31	88.95	87.39	99.14	241.85
Inputs using sym5 cDs ($j = 4$).	97.47	96.32	96.17	98.75	72.93
Sigmoid MLP(64, 32, 8, 1)					
Inputs without DWT.	50.01	41.24	18.06	82.0	7.86
Inputs using coif5 cAs ($J = 6$).	50.68	39.66	23.51	77.98	0.43
Relu MLP(64, 32, 8, 1)					
Inputs without DWT.	79.92	72.81	59.48	99.96	16.7
Inputs using sym5 cDs ($j = 4$).	97.48	96.66	96.57	98.38	7.16
LS					
Inputs without DWT.	37.99	13.65	72.22	4.34	10.57
Inputs using sym5 cDs ($j = 6$).	49.62	0	99.89	0	0.35
Random Forests					
Inputs without DWT.	97.82	97.25	97.17	98.45	9.42
Inputs using db1 cAs ($J = 4$).	98.08	97.49	97.41	98.73	1.16
CNN					
Inputs without DWT.	97.68	99.94	97.68	95.48	54.17
Inputs using sym5 cDs ($j = 1$).	99.13	99.16	99.13	99.09	22.93
Naive Bayes					
Inputs without DWT.	94.75	90.81	92.42	99.67	8.8
Inputs using bior2.4 cDs ($j = 6$).	95.95	93.37	92.85	98.99	0.35
SVM					
Inputs without DWT.	93.76	88.99	87.41	100	36.31
Inputs using sym5 cDs ($j = 5$).	94.96	90.93	89.84	99.99	1.52

coefficients. For instance, if $j = 6$ is written, it means that the best results were obtained using the cDs from the sixth level of decomposition. We executed the models in a computer with an Intel Core i7-7700 HQ CPU, 16.0 GB of RAM, Windows 10 operative system of 64 bits, and a NVIDIA GeForce GTX 1060 graphics card.

In order to statistically demonstrate that the results obtained by DWT were better than those obtained from using the binned light curves without MRA, we performed hypothesis tests. We applied the Welch's t-test to reject the possibility that the results had equal means. The results from our hypothesis tests are presented in Table 10; where we highlighted with bold letters those cases in which the null hypothesis that both means are equal is not rejected. As it can be seen, in most cases, the null hypothesis that both results are equal is rejected (when the p-value is lower than 0.05). This confirms that the results obtained with DWT are better in most cases; with some exceptions such as the Sigmoid MLP(64, 32, 8, 1) where the only amelioration is execution time for both datasets. The most remarkable improvement obtained by using DWT is execution time, which in some cases was reduced to more than a tenth of its original value. This is due to the fact that the DWT reduces the size of the input vector. However, reducing the execution time does not affect the other performance metrics, because it extracts the most relevant features from the complete signal. In fact, in general, the performance of the models was ameliorated by applying DWT to the inputs. One of the most remarkable results was obtained with the Sigmoid MLP(5, 2) tested on the Real-LC dataset, where the average accuracy was improved from 49.65% to 92.0%. Furthermore, we noticed that all the tested ML models have very different performance results, even when not using the DWT inputs. For instance, the results obtained by the different MLP configurations have significant variations in performance. A model with more neural units such, as the Sigmoid MLP(1024), has a greater execution time but also obtains better results. Even more, it is noticeable that this particular model did not obtain any improvement other than the execution time reduction from the DWT inputs. The best result of all was obtained by the CNN using DWT in the 3-median dataset, with an accuracy of 99.13% and an execution time reduction of more than a half. Since only one level of decomposition was used, the architecture used is the same as the one presented in (Shallue and Vanderburg 2018); which means that MRA did improve the

Table 10 p-values obtained from the Welch's t-tests

Dataset	Model	Accuracy	Precision	Specificity	Recall	Time
Real-LC	Sigmoid MLP(5, 2).	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
	Relu MLP(5, 2).	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
	Sigmoid MLP(1024).	0.9391	0.4999	0.3855	0.5743	< 0.05
	Relu MLP(1024).	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
	Sigmoid MLP(64, 32, 8, 1).	0.8794	0.4801	0.6697	0.6696	< 0.05
	Relu MLP(64, 32, 8, 1).	0.6357	< 0.05	0.8663	0.4457	< 0.05
	LS.	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
	Random Forests.	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
	CNN.	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
	Naive Bayes.	< 0.05	< 0.05	0.0975	< 0.05	< 0.05
	SVM.	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
3-median	Sigmoid MLP(5, 2).	< 0.05	< 0.05	< 0.05	0.069	< 0.05
	Relu MLP(5, 2).	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
	Sigmoid MLP(1024).	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
	Relu MLP(1024).	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
	Sigmoid MLP(64, 32, 8, 1).	0.0504	0.58	0.33	0.48	< 0.05
	Relu MLP(64, 32, 8, 1).	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
	LS.	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
	Random Forests.	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
	CNN.	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
	Naive Bayes.	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
	SVM.	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05

A p-value lower than 0.05 rejects the hypothesis that the results are equal performance of their model. It is also important to highlight the results obtained by the Random Forest since we were able to improve its performance; even though it gave the best results among all classifiers without DWT inputs. This was expected since the use of the DWT coefficients as features accounts for the extraction of relevant features from the light curves.

Finally, since the improvement in time is caused by the reduction of data of the DWT; we tested the effect of reducing the number of data points without performing DWT. We reduced the number of bins from 2048 to 32 (which would be the length of the inputs using DWT with six decomposition levels); then, we measured the average accuracy and time. The results are presented in Figs. 11 and 12, where it can be observed that reducing the number of bins may reduce the performance of the ML models (e.g. the performance of the MLP(5, 2) Sigmoid model decreases from 84.64 to 49.69 in the 3-median dataset); while significantly reducing the execution time. This demonstrates that, in order to reduce execution time without affecting identification performance, it is important to extract the most relevant features from the light curves, by using the DWT coefficients. Nevertheless, the binning process is still necessary because it provides a fixed length to the inputs of the ML models.

Discussion

In this survey, several ML approaches focused on the transit method for exoplanet discovery have been examined. It is noticeable that the best performing approaches for exoplanet detection and identification are related to CNNs, notably the one presented in (Shallue and Vanderburg 2018) (which presents the discovery and validation of two new exoplanets, and which is used as a template by other models), and (Pearson et al. 2018). This could be due to their capability to learn their own features, without having to depend on human extracted features. A common problem in the presented approaches is that astrophysical and instrumental phenomena within the light curves affect the performance of the algorithms. Some approaches presented in this survey can be used for preprocessing the light curves, thus, reducing the instrumental and astrophysical noise from the transit signals. However, even though these methods reduce the quantity of missed detections, they are not yet unerring. Furthermore, the noise within the light curves increases the probability of cataloguing non-transit signals as false positives. Another problem within different approaches is that low SNR signals have to be carefully analyzed; because they are likely to cause more false positives. Nevertheless, ignoring signals with such SNR values may lead to ignoring shallow transits that could belong to Earth-like planets.

It is noticeable (from the works discussed along this survey) that multiresolution analysis is mainly used for preprocessing purposes. Nevertheless, our results have shown that its use is beneficial for the identification step, and it is probable that it is also beneficial for the exoplanet detection step, in which MRA has not been fully exploited yet. Wavelets are adjusted in scale and position by using the translation and dilation operations, thus, they are perfect for describing a function in terms of building blocks generated from the mother wavelet ((Alarcon-Aquino et al. 2014; Alarcon-Aquino and Barria 2009)). This is important

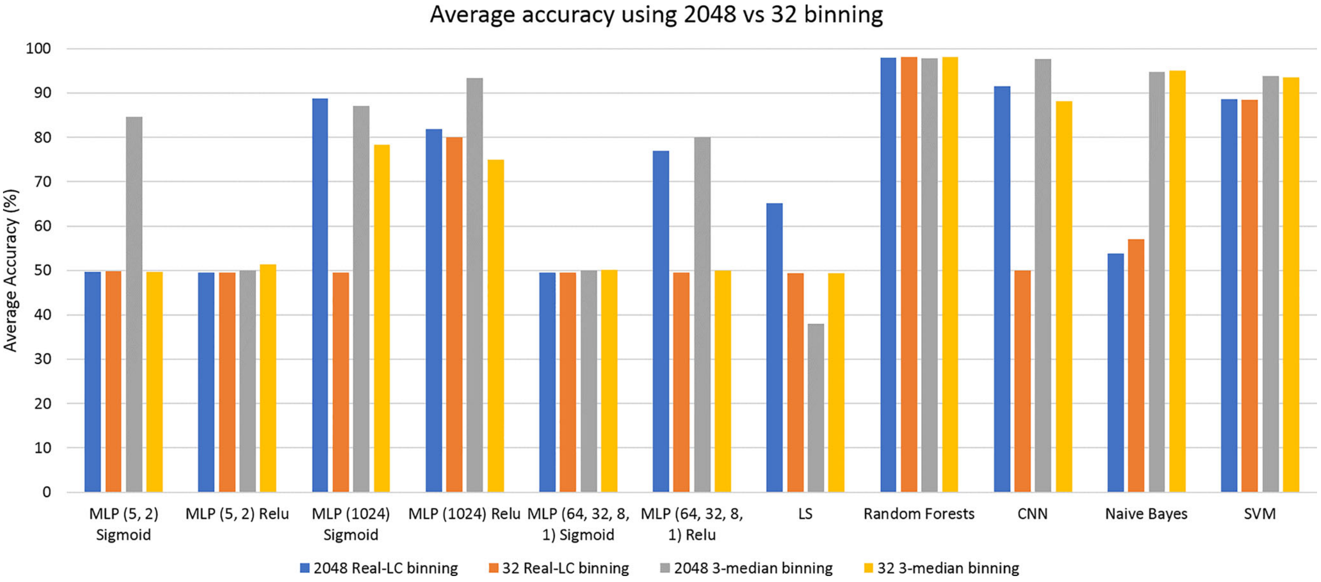

Fig. 11 Average accuracy results comparison between the 2048 and 32 binning size

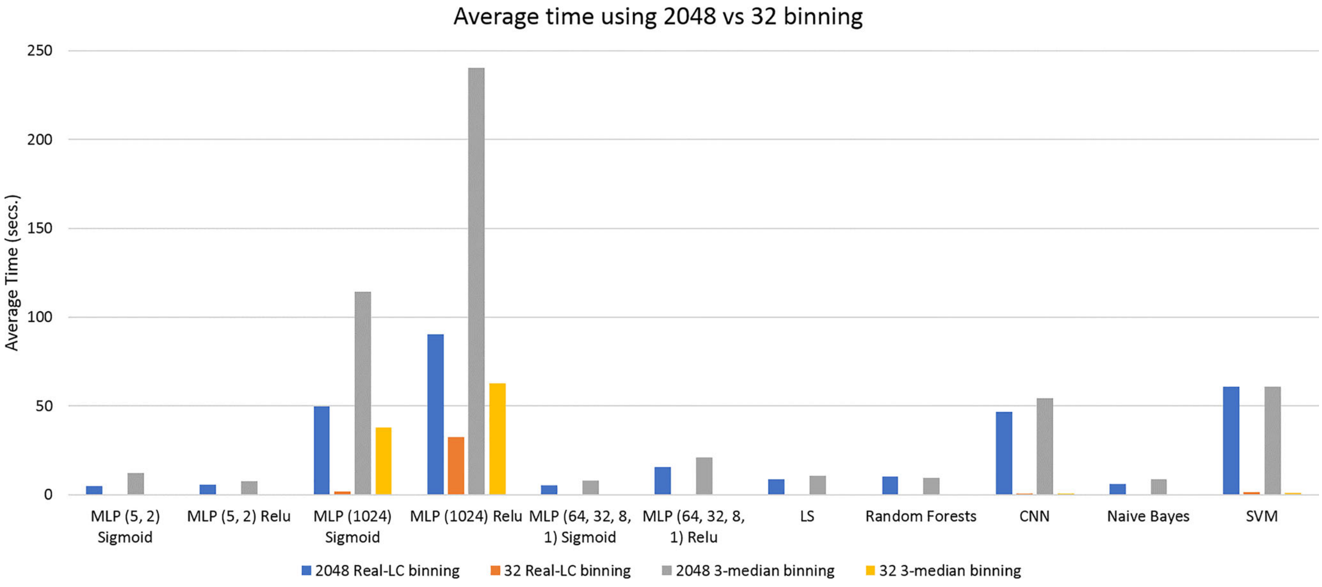

Fig. 12 Average time results comparison between the 2048 and 32 binning size

because as (Shallue and Vanderburg 2018) state, it is more likely to detect and identify new transiting planets in multi-transiting-planet systems, where different transiting signals may be mixed. The use of Fourier analysis is not suitable for this problem because, in the frequency domain, transit energy is not highly concentrated and it is masked by observational noise ((Moutou and Pont 2006)). (Moutou and Pont 2006) assert that transit detection algorithms should be sensitive enough to retrieve shallow transits, while not producing too many false alarms. (Grziwa and Pätzold 2016) state that a model independent filter is needed to correct for the variability for transit detection, such as highly variable stellar light curves, which are one of the main noise features that affect exoplanetary transit detection. They use wavelets to detect shallow transits due to their capability of analyzing the signal at different resolutions, thus evading the effects of noise. Furthermore, as stated by (Carpano et al. 2003), wavelet filters take into account the fact that noise is likely to change over the duration of the observations. For example, the wavelet filter of (Jenkins 2002) measures the dependence of noise variance in both frequency, and time by using wavelet decomposition. Moreover, the facility of wavelets to adjust to diverse non-linear functions could be used for a better feature extraction than the one obtained by the current detection and identification approaches.

In addition, (Pearson et al. 2018) state that training a neural network on wavelet components could allow one to obtain more significative pieces of the original light curve signal. This is because wavelets represent a signal as a series of components that discard the pieces that do not define the signal (such as noise), leaving the original signal intact. Another problem pointed by some solutions (such as (Petigura et al. 2013)) is the need of a system capable of detecting small Earth-like planets found on high noise light curve signals. It is stated in (Pearson et al. 2018) that the search for smaller planets depends on beating down the noise enough to detect a signal. Once more, MRA could solve this problem; e.g. by taking advantage of the wavelet dilation capacity to analyze light curve signals at different resolution levels; and to differentiate, whether the exoplanet candidate is really a transiting signal, or it is caused by noise.

Conclusions

Exoplanet research is a complex science that requires simultaneously analyzing incredibly big amounts of data, while accounting for noisy features. ML approaches can perform exhaustive tasks such as examining transiting light curves to look for exoplanet signals. This survey presented an analysis on ML approaches used for exoplanet discovery, by using the transit exoplanet detection method. While results show that there has been a great advance in this area, there are still several problems to solve. It is noticeable that noise within the light curve signals is one of the most problematic features that obstruct the full capacity of ML approaches. Noisy features can deceive AI algorithms by generating false positives or even hiding the transit signals from the detection models. Even though the current ML approaches reduce the work load for scientists dedicated to the validation of exoplanet findings, human intervention is still required (e.g. for feature extraction). In addition, weak transit signals present a great opportunity for finding Earth-like exoplanets. The ideal ML model should be capable of analyzing feeble signals. This requires a higher grade of detection and identification performance (to overcome the problem imposed by transits found in low SNR light curves). For this reasons, MRA seems to be a promising solution for detecting small planets, and vetting the detected signals.

We have proposed a model to create synthetic datasets of light curves. Using this model, we have performed experiments with two different datasets of simulated light curves, which demonstrate that MRA is capable of extracting subtle features from the light curves; while also reducing the size of the data. This helps to significantly shorten the execution time, and improve the identification performance of the ML models. Future work will be done in evaluating the performance of other MRA techniques such as Empirical Mode Decomposition (*EMD*, (Zeiler et al. 2010)), Ensemble Empirical Mode Decomposition (*EEMD*, (Huang and Wu 2008)), Stationary Wavelet Transform (*SWT*, (Nason and Silverman 1995)), as well as using the reconstructed signal in the exoplanet identification step.

Acknowledgments The authors would like to thank the Mexican National Council for Science and Technology (CONACyT) and the Universidad de las Americas Puebla (UDLAP) for their support through the doctoral scholarship program. Also, the authors would like to thank Kyle A. Pearson for his valuable feedback regarding the light curve preprocessing step. This paper includes data collected by the Kepler mission and obtained from the MAST data archive at the Space Telescope Science Institute (STScI). Funding for the Kepler mission is provided by the NASA Science Mission Directorate. STScI is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS 5–26555.

References

Aigrain S, Favata F (2002) Bayesian detection of planetary transits. a modified version of the Gregory-Loredo method for bayesian periodic signal detection. Astron Astrophys 395:625–636. https://doi.org/10.1051/0004-6361:20021290

Aigrain S, Parviainen H, Roberts S, Reece S, Evans T (2017) Robust, open-source removal of systematics in Kepler data. Mon Not R Astron Soc 471:759–769. https://doi.org/10.1093/mnras/stx1422

Akeson RL et al (2013) The NASA Exoplanet Archive: Data and Tools for Exoplanet Research. Publ Astron Soc Pac 125:989. https://doi.org/10.1086/672273

Alarcon-Aquino V, Barria J (2009) Change detection in time series using the maximal overlap discrete wavelet transform. Lat Am Appl Res 39:145–152

Alarcon-Aquino V, Barria JA (2001) Anomaly detection in communication networks using wavelets. In: IEE Proceedings - Communications, 148:355–362, https://doi.org/10.1049/ip-com:20010659

Alarcon-Aquino V, Barria JA (2006) Multiresolution fir neural-network-based learning algorithm applied to network traffic prediction. IEEE T Syst Man Cy C 36:208–220. https://doi.org/10.1109/TSMCC.2004.843217

Alarcon-Aquino V, Ramirez-Cortes J, Gomez-Gil P, Starostenko O, Garcia-Gonzalez Y (2014) Network intrusion detection using self-recurrent wavelet neural network with multidimensional radial wavelons. Inf Technol Control 43:347–358. https://doi.org/10.5755/j01.itc.43.4.4626

Anglada-Escudé et al (2016) A terrestrial planet candidate in a temperate orbit around proxima centauri. Nature 536:437–440. https://doi.org/10.1038/nature19106

Ansdell M et al (2018) Scientific Domain Knowledge Improves Exoplanet Transit Classification with Deep Learning. Astrophys J Lett 869:L7. https://doi.org/10.3847/2041-8213/aaf23b

Armstrong D et al (2015) K2 variable catalogue II: Machine learning classification of variable stars and eclipsing binaries in K2 fields 0-4. Mon Not R Astron Soc 456:2260–2272. https://doi.org/10.1093/mnras/stv2836

Armstrong D et al (2018) Automatic vetting of planet candidates from ground-based surveys: Machine learning with NGTS. Mon Not R Astron Soc 478:4225–4237. https://doi.org/10.1093/MNRAS/STY1313

Armstrong DJ, Pollacco D, Santerne A (2016) Transit shapes and self-organizing maps as a tool for ranking planetary candidates: application to Kepler and K2. Mon Not R Astron Soc 465:2634–2642. https://doi.org/10.1093/mnras/stw2881

Auvergne M et al (2009) The CoRoT satellite in flight: description and performance. Astronomy & Astrophysics 506:411–424. https://doi.org/10.1051/0004-6361/200810860

Baluev R (2018) Planetpack3: A radial-velocity and transit analysis tool for exoplanets. Astronomy and Computing 25:221–229. https://doi.org/https://doi.org/10.1016/j.ascom.2018.10.005

Baluev RV (2013) Detecting multiple periodicities in observational data with the multifrequency periodogram—ii. frequency decomposer, a parallelized time-series analysis algorithm. Astronomy and Computing 3-4:50–57. https://doi.org/https://doi.org/10.1016/j.ascom.2013.11.003

Barclay T et al (2013) A sub-Mercury-sized exoplanet. Nature 494:452–454. https://doi.org/10.1038/nature11914

Basri G, Borucki WJ, Koch D (2005) The Kepler Mission: A wide-field transit search for terrestrial planets. New Astronomy Reviews 49:478–485. https://doi.org/https://doi.org/10.1016/j.newar.2005.08.026

Beck T et al (2017) The CHEOPS characterising exoplanet satellite mission: telescope optical design, development status and main technical and programmatic challenges, vol 10562

Beuzit JL et al (2019) SPHERE: The exoplanet imager for the Very Large Telescope. Astronomy and Astrophysics 631:A155. https://doi.org/10.1051/0004-6361/201935251. arXiv:1902.04080

Bonse MJ, Quanz SP, Amara A (2018) Wavelet based speckle suppression for exoplanet imaging - Application of a de-noising technique in the time domain. arXiv:1804.05063

Borucki WJ et al (2010) Kepler planet-detection mission: Introduction and first results. Science 327:977–980. https://doi.org/10.1126/science.1185402

Bravo JP, Roque S, Estrela R, Leão I. C., De Medeiros JR (2014) Wavelets: a powerful tool for studying rotation, activity, and pulsation in Kepler and CoRoT stellar light curves. Astronomy & Astrophysics 568:A34. https://doi.org/10.1051/0004-6361/201323032

Burrows A et al (2001) The theory of brown dwarfs and extrasolar giant planets. Reviews of Modern Physics 73:719–765. https://doi.org/10.1103/RevModPhys.73.719

Campbell B et al (1988) A search for substellar companions to solar-type stars. Astrophysical Journal 331:902–921. https://doi.org/10.1086/166608

Carpano S, Aigrain S, Favata F (2003) Detecting planetary transits in the presence of stellar variability, optimal filtering and the use of colour information. Astronomy and Astrophysics 401:743–753. https://doi.org/10.1051/0004-6361:20030093

Carter J, Nathan Winn J (2009) Parameter estimation from time-series data with correlated errors: A wavelet-based method and its application to transit light curves. Astrophys J Lett 704:51–67. https://doi.org/10.1088/0004-637X/704/1/51

Catanzarite JH (2015) Autovetter planet candidate catalog for Q1-Q17 data release 24 Astronomy & Astrophysics

Charbonneau D et al (2000) Detection of planetary transits a cross a sun-like star. Astrophys J Lett 529:L45–L48. https://doi.org/10.1086/312457

Chauvin G et al (2004) A giant planet candidate near a young brown dwarf. direct VLT/NACO observations using ir wave-front sensing. Astronomy and Astrophysics 425:L29–L32. https://doi.org/10.1051/0004-6361:200400056

Chintarungruangchai P, Jiang I.-G. (2019) Detecting exoplanet transits through machine-learning techniques with convolutional neural networks. Publ Astron Soc Pac 131:064502. https://doi.org/10.1088/1538-3873/ab13d3

Cochran WD et al (2002) A Planetary Companion to the Binary Star Gamma Cephei. In: AAS/Division for Planetary Sciences Meeting Abstracts #34 p. infopages 916. volume info volume 34 of info series Bulletin of the American Astronomical Society

Coughlin JL (2017) Planet Detection Metrics: Robovetter Completeness and Effectiveness for Data Release 25 info type Technical Report NASA

Coughlin JL et al (2016) Planetary candidates observed by Kepler. VII. the first fully uniform catalog based on the entire 48-month data set Q1–Q17 DR24. Astrophys J Lett Supplement Series 224:12. https://doi.org/10.3847/0067-0049/224/1/12

Dattilo A et al (2019) Identifying Exoplanets with Deep Learning. II. Two New Super-Earths Uncovered by a Neural Network in K2 Data. The Astronomical Journal 157:169. https://doi.org/10.3847/1538-3881/ab0e12

Daubechies I (1992) Ten Lectures on Wavelets Society for Industrial and Applied Mathematics

Emmanoulopoulos D et al (2013) Generating artificial light curves: revisited and updated. Mon Not R Astron Soc 433:907–927. https://doi.org/10.1093/mnras/stt764

von Essen C et al (2018) Kepler Object of Interest Network I. First results combining ground and space-based observations of Kepler systems with transit timing variations. Astronomy & Astrophysics 615:1–16. https://doi.org/10.1051/0004-6361/201732483

Fleck B (1995) The soho mission. In: Benz A. O., Krüger Eds A. (eds). Coronal Magnetic Energy Releases. Lecture Notes in Physics, Springer, Berlin, Heidelberg. 444:233–244

Foreman-Mackey et al (2015) A systematic search for transiting planets in the K2 data. Astrophys J Lett 806:215. https://doi.org/10.1088/0004-637x/806/2/215

Freudenthal J et al (2018) Kepler Object of Interest Network. II. Photodynamical modelling of Kepler-9 over 8 years of transit observations. Astronomy & Astrophysics 618:A41. https://doi.org/10.1051/0004-6361/201833436

Gardner JP et al (2006) The James Webb Space Telescope. Space Science Reviews 123:485–606. https://doi.org/10.1007/s11214-006-8315-7

Grziwa S, Pätzold M. (2016) Wavelet-based filter methods to detect small transiting planets in stellar light curves. arXiv:1607.08417

Grziwa S, Pätzold M., Carone L (2012) The needle in the haystack: Searching for transiting extrasolar planets in CoRoT stellar light curves. Mon Not R Astron Soc 420:1045–1052. https://doi.org/10.1111/j.1365-2966.2011.19970.x

Hartman JD et al (2015) HATS-6b: A Warm Saturn Transiting an Early M Dwarf Star, and a Set of Empirical Relations for Characterizing K and M Dwarf Planet Hosts. The Astronomical Journal 149:166. https://doi.org/10.1088/0004-6256/149/5/166

He X, Niyogi P (2004) Locality preserving projections, Eds. Advances in Neural Information Processing Systems 16 pp. info pages 153–160 publisher MIT Press S. Thrun, L. K. Saul, B. Schölkopf (eds)

Henry GW, et al. (2000) A transiting ”51 peg-like” planet. Astrophys J Lett 529:L41–L44. https://doi.org/10.1086/312458

Howell SB et al (2014) The K2 Mission: Characterization and early results. Publ Astron Soc Pac 126:398–408. https://doi.org/10.1086/676406

Huang NE, Wu Z (2008) A review on Hilbert-Huang transform: Method and its applications to geophysical studies. Reviews of Geophysics 46:1–23. https://doi.org/10.1029/2007RG000228

Japkowicz N, Shah M (2011) Evaluating learning algorithms: A classification perspective

Jenkins J et al (2015) Discovery and validation of Kepler-452b: A 1.6-Re super earth exoplanet in the habitable zone of a G2 star. The Astronomical Journal 150:1–19. https://doi.org/10.1088/0004-6256/150/2/56

Jenkins JM (2002) The impact of solar-like variability on the detectability of transiting terrestrial planets. Astrophys J Lett 575:493–505. https://doi.org/10.1086/341136

Jenkins JM et al (2010) Overview of the Kepler science processing pipeline. Astrophys J Lett 713:L87–L91. https://doi.org/10.1088/2041-8205/713/2/l87

Khan MS, Stewart Jenkins J, Yoma N (2017) Discovering new worlds: a review of signal processing methods for detecting exoplanets from astronomical radial velocity data. IEEE Signal Processing Magazine 34:104–115. https://doi.org/10.1109/MSP.2016.2617293

Kingma D, Ba J (2014) Adam: A method for stochastic optimization, International Conference on Learning Representations

Koch DG et al (2010) Kepler Mission design, realized photometric performance, and early science. Astrophys J Lett Letters 713:L79–L86. https://doi.org/10.1088/2041-8205/713/2/L79. arXiv:1001.0268

Kovacs G (2017) Synergies between exoplanet surveys and variable star research. EPJ Web of Conferences 152:01005. https://doi.org/10.1051/epjconf/201715201005

Kovacs G, Bakos G, W Noyes R (2005) A trend filtering algorithm for wide field variability surveys. Mon Not R Astron Soc 356:557–567. https://doi.org/10.1111/j.1365-2966.2004.08479.x

Kovács G., Zucker S, Mazeh T (2002) A box-fitting algorithm in the search for periodic transits. Astronomy and Astrophysics 391:369–377. https://doi.org/10.1051/0004-6361:20020802

Kreidberg L (2015) batman: BAsic transit model cAlculatioN in python. Publ Astron Soc Pac 127:1161–1165. https://doi.org/10.1086/683602

Latham DW et al (1989) The unseen companion of HD114762 - A probable brown dwarf. Nature 339:38–40. https://doi.org/10.1038/339038a0

Lenzen R et al (1998) CONICA: The high-resolution near-infrared camera for the ESO VLT. In: Proc.SPIE pp. info pages 3354–3354 – 9 3354, https://doi.org/10.1117/12.317287

Males JR et al (2014) Direct imaging of exoplanets in the habitable zone with adaptive optics. In: Adaptive Optics Systems IV pp. info pages 1–13 9148 of info series Society of Photo-Optical Instrumentation Engineers SPIE Conference Series, https://doi.org/10.1117/12.2057135

Mandel K, Agol E (2002) Analytic light curves for planetary transit searches. Astrophys J Lett 580:L171–L175. https://doi.org/10.1086/345520

Masciadri E, Raga A (2004) Exoplanet recognition using a wavelet analysis technique. Astrophys J Lett 611:137–140. https://doi.org/10.1086/423984

Mathur S et al (2017) Revised stellar properties of Kepler Targets for the Q1-17 DR 25 transit detection run. Astrophys J Lett Supplement Series 229:30. https://doi.org/10.3847/1538-4365/229/2/30

Mayor M, Queloz D (1995) A Jupiter-mass companion to a solar-type star. info journal Nature 378:355–359. https://doi.org/10.1038/378355a0

McCauliff SD et al (2015) Automatic classification of Kepler planetary transit candidates. Astrophys J Lett 806:6. https://doi.org/10.1088/0004-637x/806/1/6

Morton TD et al (2016) False Positive Probabilities for all Kepler Objects of Interest: 1284 Newly Validated Planets and 428 Likely False Positives. Astrophys J Lett 822:86. https://doi.org/10.3847/0004-637X/822/2/86

Moutou C, Pont F (2006) Detection and characterization of extrasolar planets: the transit method. Ecole de Goutelas 28:55–79

Nason GP, Silverman BW (1995) The stationary wavelet transform and some statistical applications. In: Antoniadis A., Oppenheim G. (eds) Wavelets and Statistics. Lecture Notes in Statistics, Springer, New York, NY. 103:281–299

Nun I et al (2014) Supervised Detection of Anomalous Light Curves in Massive Astronomical Catalogs. Astrophys J Lett 793:23. https://doi.org/10.1088/0004-637X/793/1/23

P Hatzes A (2014) The role of space telescopes in the characterization of transiting exoplanets. journal Nature 513:353–7. https://doi.org/10.1038/nature13783

Parviainen H (2015) pytransit: fast and easy exoplanet transit modelling in python. Mon Not R Astron Soc 450:3233–3238. https://doi.org/10.1093/mnras/stv894

Pasquale BA et al (2017) Optical Design of the WFIRST Phase-A Wide Field Instrument. In: Optical Design and Fabrication 2017 Freeform, IODC, OFT ITh1B.2 Optical Society of America, https://doi.org/10.1364/IODC.2017.ITh1B.2

Pearson KA, Palafox L, Griffith CA (2018) Searching for exoplanets using artificial intelligence. Mon Not R Astron Soc 474:478–491. https://doi.org/10.1093/mnras/stx2761

Petigura EA, Marcy GW, Howard AW (2013) A plateau in the planet population below twice the size of earth. Astrophys J Lett 770:69. https://doi.org/10.1088/0004-637x/770/1/69

Pollacco D et al (2006) The wasp project and the super wasp cameras. Publ Astron Soc Pac 118:1407–1418. https://doi.org/10.1086/508556

Rauer H et al (2014) The PLATO 2.0 mission. Experimental Astronomy 38:249–330. https://doi.org/10.1007/s10686-014-9383-4

Ricker GR et al (2015) Transiting Exoplanet Survey Satellite TESS. Journal of Astronomical Telescopes, Instruments, and Systems 1:014003. https://doi.org/10.1117/1.JATIS.1.1.014003

Rodriguez JE, et al. (2018) A Compact Multi-planet System with a Significantly Misaligned Ultra Short Period Planet. The Astronomical Journal 156:245. https://doi.org/10.3847/1538-3881/aae530

Rousset G et al (2000) Status of the VLT Nasmyth adaptive optics system NAOS. Proc.SPIE 4007:4007–10. https://doi.org/10.1117/12.390304

Sanders G (2013) The Thirty Meter Telescope TMT: An International Observatory. Journal of Astrophysics and Astronomy 34:81–86. https://doi.org/10.1007/s12036-013-9169-5

Schanche N et al (2019) Machine-learning approaches to exoplanet transit detection and candidate validation in wide-field ground-based surveys. Mon Not R Astron Soc 483:5534–5547. https://doi.org/10.1093/mnras/sty3146

Schwamb ME et al (2013) Planet hunters: A transiting circumbinary planet in a quadruple star system. Astrophys J Lett 768:127. https://doi.org/10.1088/0004-637x/768/2/127

Seager S, Bains W (2015) The search for signs of life on exoplanets at the interface of chemistry and planetary science. Science Advances 1:e1500047–e1500047. https://doi.org/10.1126/sciadv.1500047

Seager S, Mallén-Ornelas G (2003) A unique solution of planet and star parameters from an extrasolar planet transit light curve. Astrophys J Lett 585:1038–1055. https://doi.org/10.1086/346105

Shallue CJ, Vanderburg A (2018) Identifying Exoplanets with Deep Learning: A Five-planet Resonant Chain around Kepler-80 and an Eighth Planet around Kepler-90. The Astronomical Journal 155:94. https://doi.org/10.3847/1538-3881/aa9e09

Smith JC et al (2012) Kepler Presearch Data Conditioning II - a bayesian approach to systematic error correction. Publ Astron Soc Pac 124:1000–1014. https://doi.org/10.1086/667697

Stumpe MC et al (2014) Multiscale systematic error correction via wavelet-based bandsplitting in Kepler Data. Publ Astron Soc Pac 126:100–114. https://doi.org/10.1086/674989

Tamuz O, Mazeh T, Zucker S (2005) Correcting systematic effects in a large set of photometric light curves. Mon Not R Astron Soc 356:1466–1470. https://doi.org/10.1111/j.1365-2966.2004.08585.x

Thompson SE et al (2015) A machine learning technique to identify transit shaped signals. Astrophys J Lett 812:46. https://doi.org/10.1088/0004-637x/812/1/46

Tingley B (2003) Improvements to existing transit detection algorithms and their comparison. Astronomy and Astrophysics 408:L5–L7. https://doi.org/10.1051/0004-6361:20031138

Treu T, Marshall PJ, Clowe D (2012) Resource Letter GL-1: Gravitational Lensing. American Journal of Physics 80:753–763. https://doi.org/10.1119/1.4726204. arXiv:1206.0791

Vanderburg A, Johnson JA (2014) A Technique for Extracting Highly Precise Photometry for the Two-Wheeled Kepler Mission. Publ Astron Soc Pac 126:948. https://doi.org/10.1086/678764

Veitch D (2005) Wavelet Neural Networks and their application in the study of dynamical systems Master's thesis University of York

Way MJ et al (2012) Advances in Machine Learning and Data Mining for Astronomy 1st ed. info publisher Chapman & Hall/CRC

Werner MW et al (2004) The Spitzer Space Telescope Mission. Astrophys J Lett Supplement Series 154:1–9. https://doi.org/10.1086/422992

Wheatley PJ et al (2018) The Next Generation Transit Survey NGTS. Mon Not R Astron Soc 475:4476–4493. https://doi.org/10.1093/mnras/stx2836

Wolszczan A, Frail D (1992) A planetary system around the millisecond pulsar PSR1257 + 12. Nature 355:145–147. https://doi.org/10.1038/355145a0

Wootten A, Thompson AR (2009) The atacama large millimeter/submillimeter array. In: Proceedings of the IEEE, vol 97, pp 1463–1471, https://doi.org/10.1109/JPROC.2009.2020572

Yaqoob T (2011) Exoplanets and Alien Solar Systems New Earth Labs

Yu L et al (2019) Identifying Exoplanets with Deep Learning III: Automated Triage and Vetting of TESS Candidates. arXiv:1904.02726

Zapatero Osorio MR et al (2000) Discovery of young, isolated planetary mass objects in the σ orionis star cluster. Science 290:103–107. https://doi.org/10.1126/science.290.5489.103

Zeiler A et al (2010) Empirical Mode Decomposition - an introduction. In: Proceedings of the International Joint Conference on Neural Networks 1–8, https://doi.org/10.1109/IJCNN.2010.5596829

Zingales T et al (2018) The ARIEL mission reference sample. Experimental Astronomy 46:67–100. https://doi.org/10.1007/s10686-018-9572-7

Zucker S, Giryes R (2018) Shallow Transits-Deep Learning. I Feasibility Study of Deep Learning to Detect Periodic Transits of Exoplanets. The Astronomical Journal 155:147. https://doi.org/10.3847/1538-3881/aaae05