CLEI electronic journal, Volume 22, Number 3, Paper 3, December 2019

Classical Machine Learning Techniques in the Search of Extrasolar Planets

Margarita Bugueño*
Univ. Técnica Federico Santa María, Dept. de Informática
Santiago, Chile, 8940000
margarita.bugueno@usm.cl

and

Francisco Mena*
Univ. Técnica Federico Santa María, Dept. de Informática
Santiago, Chile, 8940000
francisco.menat@usm.cl

and

Mauricio Araya
Univ. Técnica Federico Santa María, Dept. de Electrónica
Valparaíso, Chile, 2340000
mauricio.araya@usm.cl

Abstract

The field of astronomical data analysis has experienced an important paradigm shift in the recent years. The automation of certain analysis procedures is no longer a desirable feature for reducing the human effort, but a must have asset for coping with the extremely large datasets that new instrumentation technologies are producing. In particular, the detection of transit planets — bodies that move across the face of another body — is an ideal setup for intelligent automation. Knowing if the variation within a light curve is evidence of a planet, requires applying advanced pattern recognition methods to a very large number of candidate stars. Here we present a supervised learning approach to refine the results produced by a case-by-case analysis of light-curves, harnessing the generalization power of machine learning techniques to predict the currently unclassified light-curves. The method uses feature engineering to find a suitable representation for classification, and different performance criteria to evaluate them and decide. Our results show that this automatic technique can help to speed up the very time-consuming manual process that is currently done by expert scientists.

Keywords: machine learning, exoplanet detection, feature engineering, light-curve

1 Introduction

Planets orbiting stars outside our solar systems are called extra-solar planets or exoplanets. Detecting these planets is a challenging problem, because they only emit or reflect very dim magnitudes of light compared to their host stars, and they are very near to them compared to the observation distance.

The first exoplanet detection was made by the astrophysicists Aleksander Wolszczan and Dale Frail in 1992 when they found three extrasolar planets orbiting the Lich pulsar [1], also called PDR1257+12. From here, this field of study began to gain scientific interest, but it was not until 1995 that it became a rather

*Both authors contributed equally to this work.

dynamic research area in which astronomers have found, to date, more than 3500 exoplanets using different novel techniques.

Several approaches have been proposed by astronomers for detecting extrasolar planets, being the fine-grained analysis of periodicities in star light-curves the most successful so far. However, the large volume of data that is being generated by modern observatories [2], including large surveys of astronomical objects, requires the use of automatized methods that can reproduce the analysis performed by astronomers to decide if the data supports the existence of an exoplanet or not. Fortunately, the advances in numerical methods, machine learning and data science in general allow us to apply algorithms and computational techniques that learn and predict from complex patterns in a reasonable frame of time.

This paper presents how to use machine learning techniques to refine the detection of exoplanets using real data from Kepler Space Telescope. Concretely, we explored different feature extraction and selection techniques applied to the light-curves to improve the training and prediction performance of supervised classification methods. To compare them, we used a few standard statistical learning criteria, yet the results motivated a hybrid selection of techniques that reduces the number of unconfirmed light curves.

The results of this work show that using only standard statistics to summarize the entire light curve has a moderate performance compared to combining them with other manually hand-crafted features extracted by experts and cross-matching the database with other catalogs. These features add information about the planet and its hosting star beyond the periodic properties of the ligh-curve. We reached an 88.3% on the harmonic mean between precision and recall (*F1 score*), meaning that approximately the 88% of the times the prediction was clean/correct and complete. For this configuration the best learned model was Random Forest. To produce our final classification we combined the identification of *Confirmed* exoplanets using Random Forest and the identification of *False Positive* (not exoplanet) using a SVM model with RBF kernel. Thus, the present work gives the classification of the Kepler Objects of Interest that have been studied by NexScI scientific staff to September of 2017, i.e. *Candidates* objects.

This paper is organized as follows. Section 2 presents a brief introduction to the methods of extrasolar planet detection which inspired this work. Section 3 discusses the dataset of Kepler Objects of Interest and the manual classification performed over this data. Section 4 introduces some machine learning methods and its corresponding metrics to use. Section 5 proposes an experimental study with empirical results on exoplanets classification. We conclude in Section 6 presenting our remarks and outlining future work.

2 Background

The study of exoplanets is a relatively new field of astronomy which started with the first confirmed detection of a very fast-orbiting giant planet [3] in 1995, named 51 Pegasi b. Since then, the advances in instrumentation and data analysis techniques have allowed the discovery of thousands of exoplanets. For example, NASA has reported that more than 3500 exoplanet has been detected[1] using different techniques. Unfortunately, in mostly all the cases the currently observatories, grounder or spatial, applied indirect methods due to the impossibility of direct detection.

2.1 Why is difficult to detect exoplanets?

A planet is an object that orbits around a star and is massive enough to clear of dust and other debris the protoplanetary disk from which it was born. The theory of extrasolar planets is under development since mid-nineteenth century and although there were some unsubstantiated claims regarding their discovery, it was not until recently that we have confirmed detections. Now, we can start to answer questions such as how common they are and how similar they are to the Solar system planets.

Protoplanetary disks are regions of gas and dust orbiting around young stars. Current theories suggest that the dust particles begin to collapse by gravity forming larger grains. If these discs survive to stellar radiation and comets or meteorites, the matter continues compacting giving way to a new planetoid. Unfortunately, most of the planets that have been discovered are relatively large compared to the dimensions of the Earth, highlighting the limitations of the detection methods used so far: the detection of exoplanets is determined by the sensitivity of the current observatories. Likewise, planets are very dim sources of reflected light compared to their hosting stars. Therefore, it is extremely difficult to detect this type of light, with only a couple of dozen exoplanets directly photographed while the majority of known exoplanets have been detected through indirect methods.

As indicated in the Table 1, the most successful detection mechanisms are:

[1] http://exoplanets.nasa.gov

Astrometry	1
Direct visual detection	44
Radial velocity	676
Transit	2951
Gravitational micro lenses	64
Radio pulses of a pulsar	6

Table 1: Number of confirmed exoplanets according to the detection method.[3]

- *Radial velocity*, which studies the speed variations of a star product of its orbiting planets, analyzing the spectral lines of this one through the Doppler effect to measure the red-shift or blue-shift. This method has been successful, but it is only effective on giant planets near its star.
- *Transit photometry*, photometric observation of the star and detection of variations in the intensity of its light when an orbiting planet passes in front of it, blocking a fraction of the starlight. This method efficiently detect high-volume planets independently of the proximity of the planet to its star.
- The others are less successful, like the gravitational micro lenses effect. This effect occurs when the gravity field of the planet and the host star deviate the light of another distant star, this requires that the three object are aligned, with the disadvantage that measurements cannot be repeated. Or the simplest: visual detection, based on photography of visible or infrared light. This method is very difficult due to the big gap between the shine of all other stars and planets on the sky background.

Fortunately, technological advances in photometry have allowed experiments like the space observatory Kepler to have sufficient sensitivity for detecting a greater range of exoplanets. To achieve this, feature extraction, classification and regression methods and models are needed.

What is Kepler? Kepler was a space observatory launched by NASA in march of 2009 with the objective of find similar planets to Earth within our galaxy neighborhood. Kepler measured the light variation of thousand of distant stars in the search of periodic planetary transit, with the same duration and change of brightness. Kepler monitored more than 100 thousand stars similar to our Sun[2].

2.2 Previous work

The closest research area that can be mentioned in terms of detection of extrasolar planets, correspond to the classification of variable stars. Which correspond to stars that vary their light intensity through time depending on their chemical composition or because something blocks the light emitted partially, but no planets are involved. The most common cases are RR Lyrae stars. The typical approach for this is extract manual specialized features from the light curve. For example, Richards et al. [4] presented a catalog of variable stars and manually extracted 32 light curve specialized features from simple statistics, like kurtosis, skewness, standard deviation, stetson, and other features based on the period and frequency analysis of a Lomb-Scargle [5] fitted model. Donalek et al. [6] also worked on classify variable stars from the Catalina Real-Time Transient Survey (CRTS) and the Kepler Mission, extracting similar features from the light curve to Richards et al. [4]. Other application using statistical descriptors is the work of Nun et al. [7] which, with the purpose of classify variable stars, detects anomalous light curves and erase them of the training phase. She used a supervised learning that determine if the combination of the data (light curve) with the label corresponds to a true combination (not *outlier*) based on probabilistic learning models.

Likewise, some other works made based on the same data that we use here but as the Kepler mission matured and the list of planet candidates grew up, the focus of the research community was on population-level studies [8, 9] (analysis of groups of data), which translated in changing attitudes toward the way the planet candidate lists were produced and vetted. While heterogeneous catalogs that rely on human judgment are an efficient way to spur follow-up observations of Kepler's planet candidates, they are not well suited for population studies, which rely on uniformity of the group. Uniform and automatically produced catalogs make it possible to characterize both the precision and recall.

Midway through the mission, efforts were made to produce catalogs of candidates for the planets automatically, relying less and less on human judgment. Meanwhile, others began to explore the use of machine learning to automatically examine Kepler's objects. For example, McCauliff et al. [10] used a Random Forest model based on features derived from Kepler pipeline statistics while Thompson et al. [11] used unsupervised machine learning to cluster Kepler light curves features extracted with similar shapes with the focus of find candidates object of interest.

[2]https://exoplanetarchive.ipac.caltech.edu/docs/KeplerMission.html

Some other works focused on the task of classify the variability of a star, had tried different approaches known as *representation learning*, one example is the presented by Mackenzie et al. [12]. This work designed an unsupervised feature learning algorithm which clusters sliding window over the light curve with a custom distant function and used this to learn a classifier instead of the hand-crafted features. Mahabal et al. [13] transformed the light curve into an image (grid) that represented the variations of magnitude through time intervals. This work also used the data from CRTS and complete the classification task using convolutional neural networks.

The recent work of Hinners et al. [14] presented different machine learning techniques and models with the objective of classify and predict features over the same data that we use in this paper. Similarly to what we propose, they extracted some statistical features from the light curve, but they were not focused on detecting if the light curve variations were indeed generated by an exoplanet or by other phenomena. They also tried a recurrent neural network from automatic feature extraction and prediction but the results were not conclusive.

In this work we tackle the exoplanet detection problem with *ad-hoc* feature extraction over the sequence in addition of automatic techniques using the Fourier transform and component analysis. Then, combining the best learned model for each class (exoplanet or false exoplanet), we complete the proposed task and give predictions of the object under study by Kepler.

3 Data

Currently, NASA Exoplanet Science Institute has shown[3] that around 65% of exoplanet discoveries (2344) have been detected thanks to the Kepler Mission. Considering that most of the discovered exoplanets have been detected through the transit method, and taking advantage of the photometric improvements of Kepler, we propose to work with the *Kepler Objects of Interest* (KOI[4]) dataset. This dataset, provided by MAST (*Mikulski Archive for Space Telescopes*) [15], is composed by 9564 records with 44 features each, including metadata and links to the actual light curves.

To obtain the data, firstly, we downloaded the metadata from the MAST platform and then downloaded the attached files. We collected 8054 FITS (Flexible Image Transport System) [16] files from the archive, where some of them contain more then one light curve, because different *KOI*s were detected for the same star. Moreover, each file contains the error associated to each measure, the time (in Julian date) when the measure was made, the raw light curve, the filtered light curve and a Mandel-Agol fit [17] of the light curve.

Every record is associated to a Kepler Object of Interest labeled as *Confirmed*, *False Positive* or *Candidate*, according to Nasa Exoplanet Science Institute[5].

- 2281 CONFIRMED: those that through extensive analysis have been confirmed as exoplanet.
- 3976 FALSE POSITIVE: those that were initially selected as candidate exoplanets but there is additional evidence that shows they are not.
- 1797 CANDIDATE: those that are still under study.

Between the reasons to catalog a candidate as a *False Positive* according to MAST, we found observation that did not match with the star position on study, for instance because the transit was from another object (non-planetary) in the background. Another possibility is that the deep of the even transit was statistically different to the deep of the odd transits, showing a binary system, i.e two stars orbiting among them. Almost all stars under study are "new" stars that has not been studied before, and only some of them (25%) has been previously studied in the search of exoplanets.

3.1 Light curves

The data that we handle in this work are the light curves of transit objects (see Figure 1). These are sequences of light intensity that vary when a transit object pass in front of the star and block some light. Some important information regarding to the light curve that we gathered are:

- The sampling rate, or $(t_i - t_{i-1})$, of our re-collected data is 0.0204 BJD – approximately half of an hour.
- The duration of all the collected measurements was four year, between the launch of the project 2454964.51 BJD (starts of the year 2009) to 2456424.00 BJD (same days in 2013).

[3] https://exoplanetarchive.ipac.caltech.edu/docs/counts_detail.html
[4] http://archive.stsci.edu/search_fields.php?mission=kepler_koi
[5] http://nexsci.caltech.edu/

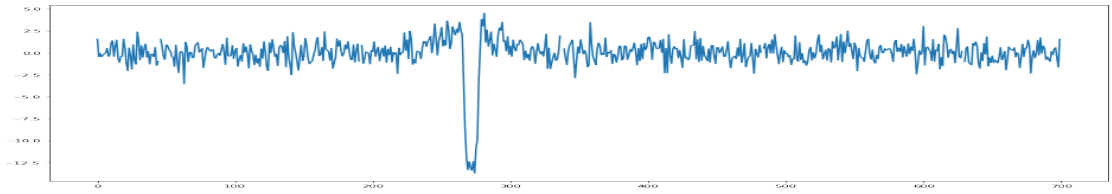


Figure 1: Sample of a light curve as a raw format, with whitened filter applied.

Even though each stored light curve (Figure 1) was about 70000 measurements, every point in the series is not generated independently [18]. As the dispersion varies through time, the series is governed by a trend and could have cycles. This fact is important because the Kepler measurements are not recorded uniformly, getting light curves with missing data. On average, the missing data is about 22.98% of the size of the fully sampled light curve, as it shown in the Appendix. This means that every light curve has approximately 55000 effective measurements, i.e. measurements with non-missing values with which the model has to learn.

From all the metadata that was available, we have selected only 10 features that we found relevant for the classification.

- *KOI count:* the numbers of identified candidates (KOI) on that system, which varies between 1 and 7.
- *Period:* the interval between consecutive planetary transits in days. Calculated as one of the best-fit parameter on a Mandel-Agol model [17].
- *Transit Depth:* the fraction of the stellar intensity lost on the minimum planetary transit.
- *Planet Radius:* the inferred radius of the Object of Interest (Earth = 1). Based on the stellar radius and a best-fit parameter on a Mandel-Agol model [17].
- *Planet Teq:* the expected temperature of equilibrium on the surface of the candidate planet in Kelvin. Calculated through the stellar temperature and the planet's radiation.
- *Stellar Teff:* is the effective stellar temperature (photosphere) in Kelvins. Based on the bright and color (luminosity) with the H-R diagram.
- *Stellar log(g):* the logarithm of the stellar surface gravity.
- *Stellar Metallicity:* is the logarithm of the relationship between Fe and H on the surface of the star, normalized by the solar relationship between Fe and H (Solar = 1).
- *Stellar Radius:* the photospheric radius of the star (Solar = 1). Based on the luminosity and temperature of the star.
- *Stellar Mass:* the inferred mass of the star (Solar = 1). Based on the mass-luminosity relation.

We used this data in order to train the models and predict correctly the object under study (Candidates).

Whitened filtering

The data used in this paper is the light curve on its raw format with a whitened filtered applied. The objective of this filter is to obtain a light curve with a constant white noise where the higher signal (high to noise) get amplified and gives a more uniform signal. Whitened filtering is a linear transformation that takes a sequence of random variable (with know covariance matrix) into a new sequence of new random variables where the covariance matrix is the identity, no correlation between variables and variance normalized. The transformation is called whitened because it changes the input data into a new data with white noise. The white noise [19] is a random signal that has the same intensity on different frequencies, which gives a spectral density of constant power. The operation realized over the light curve in Fourier domain is to divide the signal by his own spectral power density function and then return the light curve to the time domain.

$$\tilde{f}_j = \frac{f_j}{\sqrt{\Phi(f_j)}} \tag{1}$$

With f_j the j frequency component of the light curve x and the spectral power density function equal to:

$$\Phi(f_j) = S_x(f_j) \cdot S_x^*(f_j) \tag{2}$$

S_x is the Fourier transform of the light curve x and $*$ denotes the complex conjugate.

Mandel-Agol model

Within the FITS files there is also the Mandel-Agol model fitted to the light curve, which model the transit of a stratospheric planet around a stratospheric star, like an eclipse, assuming a uniform source. It requires to know the distance from the center of the planet to the center of the parent star as well as the radius of each one of the bodies. Mandel-Agol models the opacity observed on the light intensity according to the planet position. When the planet eclipse the star the opacity is maximum, when the planet orbits without eclipse the star the opacity is minimum and uniform (zero or null), yet, when the planet is close to eclipse the star the intensity is modeled as a quadratic polynomial according to [17].

4 Models and Methods

The objective of the models used is to correctly catalog the objects that are currently being investigated by NexScI (*Confirmed*) and label them as *Confirmed* or *False Positive* according to what the model learns.

4.1 Data pre-processing

As our data representation $\{(x^{(i)}), y^{(i)})\}_{i=1}^{N}$ has missing values, this is $x^{(i)} = (x_0^{(i)}, NaN, x_2^{(i)}, \ldots, x_T^{(i)})$, we had to applied techniques that could handle this properly.

- Fill in with expected value: zero. This is because it is the expected value of all the light curves and represents the stationary state in which the planet does not eclipse the star (which is almost the whole sequence).

$$\mathbb{E}[x^{(i)}] \approx 0 \quad , \forall i \in \{1, \ldots, N\} \tag{3}$$

- Fill the gaps with linear interpolation. This represents that between all the missing data there is a uniform change until the next data, that is not missing, appears. As an example for the NaN on the position $t = 1$ in input $x^{(i)}$

$$input(t) = \frac{x_2^{(i)} - x_0^{(i)}}{2 - 0} \cdot t + x_0^{(i)} \tag{4}$$

- Also we try a mix of this two by performing a *sampling* of the sequence by taking the maximum value each 3 points (considering the missing data as zeros) and finally completing the data with linear interpolation. Hoping to eliminate missing values in continuous data, a ilustration is presented:

$$\tilde{x}^{(i)} = (max(x_0^{(i)}, 0, x_2^{(i)}), \ldots, max(0, 0, x_T^{(i)})) \tag{5}$$

  And then do a linear interpolation to the *NaNs* (exactly zeros) in the processed $\tilde{x}^{(i)}$.

4.2 Feature extraction

In this work we use two different methods for feature extraction. The first one focuses on the use of manual techniques for feature extraction and construction, and the second one focuses on automatic techniques for the same propose: feature generation. Both methods are applied to the processed light curve mentioned in the previous section.

4.2.1 Manual feature extraction

We used extraction techniques specialized on time series, which in this case corresponds to measurements of light intensity through time, based on *Feature Analysis for Time Series* (*FATS*) library for Python [20]. This library was created with the purpose of extracting characteristics of astronomical time series, originally curves of light, and was previously used in [14] applying FATS over same data we use in this work. Likewise, similar features have been used on other tasks over light curves such as [4, 6] did.

Due to the performance problems experienced in making use of the library[6], i.e. execution times was indeed very high for the amount of data used, we have developed our own FATS version using some of the features present in the original implementation. Our version includes:

- *Amplitude*, defined as the difference between the maximum and the minimum value divided by 2.
- *Slope*, defined as the the slope of a linear fit to the light curve.
- *Max*, the maximum value of the sequence.
- *Mean*, the average of the sequence.
- *Median*, the median magnitude of the sequence.
- *Median Abs Dev*, defined as the median of the absolute difference between every point to the median of the sequence.
- *Min*, the minimum value of the sequence.
- *Q1*, the first quartile of the data in the sequence.
- *Q2*, the second quartile of the data in the sequence.
- *Q31*, the difference between the third and first quartiles.
- *Residual bright faint ratio*, is the rate between the residual of the fainter intensities over the brighter intensities, the mean of the sequence act as threshold.
- *Skew*, defined as a measure of the sequence asymmetry (third standardized moment).
- *Kurtosis*, defined as the fourth standardized moment of the sequence.
- *Std*, standard deviation of the sequence.

Having completed the feature extraction, we notice that only a few features were used to summarize the light-curve from a total of 55 thousand effective measurements. For this reason we decided to consider some of the metadata presented in the above sections, which could contribute with additional information of every one of the observation of the objects of interest and their light curves.

4.2.2 Automatic feature extraction

Alternatively for the automatic feature extraction techniques we use unsupervised learning methods, where the objective is to find intrinsic patterns among all the data independently of the task, in this case the exoplanet detection. The first method is the well-know Principal Component Analysis (**PCA**) [21]. This algorithm is a linear method that projects the data into a lower dimensional space, i.e. transforms the data space from the original dimensions (the length of the sequence) into a new space of lower dimensionality defined by the vector of higher variances. PCA is know as one of the best algorithm for dimensionality reduction and has been used for several applications obtaining particularly good results on time series [22–24]. Besides its great efficiency when dealing with high dimensional data, PCA can benefit from specific optimizations over linear algebra methods that are present on several libraries.

A second method is **FastICA** (*Fast algorithm for Independent Component Analysis*) [25, 26], an efficient iterative algorithm that finds statistically independent components of the data, in contrast to the uncorrelated ones used by PCA. The algorithm is focused on the signal abstraction, since it tries to detect the independently sources that, mixed, produce the observed data.

These two automatic methods were tested using the Discrete Fourier Transform (DFT) [27, 28] applied to the light curve. The DFT transforms the data from the time domain in which the measurements were obtained, to the frequency domain where the signal was generated. This method is designed to analyze periodic signals, which is exactly the case of transit light curves. The potential of Fourier lies in that it can decompose a complex signal into a subset of unique frequencies without knowing when it was produced. The objective is to recognize patterns in the characteristics that represent the frequency domain of the data.

[6] https://github.com/isadoranun/FATS

4.3 Learning Models

The K-nearest neighbors model (**k-NN**) is a popular yet simple approach based on memory (i.e., non-parametric). It remembers all the training data and uses a k number of nearest samples of the data to predict a class [29]. This algorithm classifies based on the voted majority among his k nearest neighbors, it used a distance metric to measure closeness, in our case Minkowski's [7]. Despite its simplicity, it shows a very good performance in several problems [30].

The second model is the regularized logistic regression: a variant of the logistic regression proposed in [31] that classifies based on a probabilistic (logistic) binary model in which a linear boundary is defined among the classes based on the probability of belonging to each class. The regularization is used to penalize the parameters to avoid *overfitting*, i.e., learning more complex patterns than the ones present original phenomenon for the sake of reducing the error.

Another linear algorithm, namely the Support Vector Machine (SVM) [32] is a margin-based model which also defines a linear boundary among the classes and tries to find the best separation hyperplane that divide them. SVM uses a subset of the data to fit the model: only those that are closely enough to the boundary are remembered and they are called support vectors. We use the regularized version of SVM, c-SVM, that penalizes the error on the training data, similarly to what was used for logistic regression, producing better generalization results. For both models the l_2 norm was selected (ridge regression). For the c-SVM we use a Gaussian kernel as the *Radial Basis Function* (RBF) [24] because it produces a more flexible decision boundary. Indeed, it computes the operations in a higher dimensional space, translating the problem to a non linear decision space. This technique brings improvements when the data is not linear separable.

At last, we use the Random Forest classifier [33, 34], this is an ensemble architecture in which many models (in this case decision trees) are trained over different samples of the data (bootstrap: allowing to repeat). Every model is trained by selecting some features randomly and dividing the samples according to this features. The predictions of the ensemble corresponds to the majority class among all the models. Thanks to allowing more than one depth to the decision tree in the ensemble achieve non-linear boundaries.

As an alternative to these models and feature extraction techniques, we use state-of-the-art recurrent neural network models, specifically the LSTM (Long Short Term Memory) [35] and GRU (Gated Recurrent Unit) [36], which are specifically designed for time series. The recurrent neural network considers that a sequence has a local dependency that affects the output, so it takes the dependency into account. These models are designed to cope with very large sequences, because these *gates* can detect pattern while forgetting samples that are useless and keep those that are not. Given the difficulty of training this network with the length of the sequence that we used, the light curve was transformed into a sequence of statistics by windows. In this case we used: maximum, minimum, mean, standard deviation and the third moment.

4.4 Unbalanced data

With the purpose of handling the unbalance data among the classes, we proposed to used two techniques. One consist in the subsample the majority class, in our case it is the *Confirmed* class, and with this getting two sampled sets of similar sizes as training set, known as undersampling technique [37]. Nevertheless, the other technique directly used the unbalanced data in training since it admits class weights in the objective function [38]. In this way, the minority class has more impact on it. This can only be used on Logistic Regression, SVM and Random Forest, and has the following formulation:

$$\mathcal{L} = \sum_{i}^{N} w(y_i) \cdot \ell(y_i, \hat{y}_i) \tag{6}$$

Where $w(y_i)$ is the weight of the class y_i and ℓ is the loss function of the modeling.

4.5 Metrics

The exoplanet problem is an instance of unbalanced binary classification. Therefore we need to select quality measures beyond the classical accuracy, namely *precision*, *recall* (by class) and *F1 score*, where the last one summaries *precision* and *recall* metrics in just one value over all the data (both classes).

- **Precision**
  Rate between the objects correctly labeled as one class over the sum of all the objects labeled as that class. In other words, is the ability of the model to label one class A only when the object effectively was from that class.

$$P = \frac{T_p}{T_p + F_p} \tag{7}$$

[7] http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.DistanceMetric

- **Recall**
  Rate between the objects correctly labeled as one class over the sum of all the objects effectively from that class. In other words, is the ability of the model to include all the object that effectively are from one class A.
$$R = \frac{T_p}{T_p + F_n} \tag{8}$$

- **F1 score**
  This is defined as the harmonic mean between the two measures previously mentioned, being high when both are high. Note that the relative contribution of precision and recall to the F1 score are equal. Therefore, this is a good quality measure of a model:

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{9}$$

With T_p as the true positive, F_p as the false positive and F_n as the false negative of class predictions. All these metrics reach their best values at 1 and worst at 0.

5 Experiments and results

Due to the large amount of data that was processed, it was necessary to use a cluster provided by ChiVO [2] (Chilean Virtual Observatory) in which 6257 labeled data, corresponding to 121 GB, and 1797 not labeled data (Candidates), corresponding to another 33 GB, were downloaded. Thus, following the standard set separation for machine learning, we used approximately 4000, 1000 and 1000 registers grouped as training, validation and testing sets respectively (64/18/18%). The validation set was used to tune structural hyper-parameters of the different algorithms while testing set was used to compare the best models to simulate how these will behave on future data, i.e. over Candidate class which corresponds to the unknown label objects.

Regarding to the hyper-parameters of the trained algorithms, it was necessary to define:

- k-NN: Number of neighbors k.
- SVM: C, inverse of regularization strength. Smaller values specify stronger regularization.
- Logistic Regression: C, same as SVM.
- Random Forest: The maximum depth of all the Decision Trees.

Note that the search and selection of such hyper-parameters was not expensive in computational terms because it was performed on the representations of features extracted from the techniques already discussed in previous section, which are much smaller than the amount of data ($d \ll n$, where n indicates the data size and d the dimensionality of these).

For automatic feature extraction techniques, only a few fixed dimensions were experimented due to the computational cost, **5**, 10, 15, 20, 25, 50 for ICA and **5**, 10, 25, 55, 100 for PCA. The Table 2 shows how the performance of the best model varies according to the dimensionality, it is evident that as the dimensionality increases (characteristics), the error does too. It can indicate a possible excess of information, i.e. a few attributes suffices to define a predictive model that can perform with a good generalization.

The use of a discrete Fourier transform seems to be a crucial procedure when extracting features automatically. If we apply the learning models to the extracted features on the raw data representation (sequence of intensity of light), the error turned out equally to a random labeling (i.e., all the examples as *False Positive* class 0.486, while 0.200 for *Confirmed* class in terms of *F1 score*).

Surprisingly enough, completing missing data with zeros produces a consistent improvement of ~ 0.1 in the *F1 score*, while linear interpolation and *sampling* produced worse results. We think that is because there is only a few missing values between contiguous measurements. When the *sampling* was performed, the total of the missing data (23% of the original sequence), was reduced only to 20%. Also there is the simplicity of filling in with the expected value (zero) makes it possible to avoid adding noise or badly generated synthetic data.

In addition to extracting features of the light curve manually, we worked directly with the MAST metadata features as we mentioned in previous section, results are shown in Table 3. First, the metadata corresponding to the potential exoplanet under study (KOI - *kepler object of interest*) was used. This contains the orbit period, the transit depth, the planet radius, the equilibrium temperature of the planet and the number of KOI under study in such system. This approach proved to be better than the techniques that

d	ICA
5	**0.711**
10	0.709
15	0.709
20	0.686
25	0.679
50	0.675

d	PCA
5	**0.713**
10	0.701
25	0.701
55	0.699
100	0.702
255	0.689

Table 2: *F1 score* in function of dimensionality, d, of the best classifier: Random Forest.

	Learners			
	k-NN	*Logistic Regression*	*SVM RBF*	*Random Forest*
Fourier + PCA	0.679	0.493	0.486	**0.713**
Fourier + ICA	0.679	0.493	0.486	**0.711**
OwnFATS	**0.666**	0.583	0.575	0.658
Planet metadata	0.825	0.848	0.848	**0.870**
Stellar metadata	0.766	0.718	0.751	**0.766**
OwnFATS + stellar & planet metadata	0.844	0.864	0.876	**0.883**

Table 3: *F1 score* on the classification of different models (learners) over the test set on the different representations generated.

faced the raw data. Also, we included the metadata of the hosting star such as the effective temperature, the metallicity, the gravity, the radius and its mass. When we used these features, it was possible to notice an important improvement, because we fed the classification algorithms with more relevant information than using only features extracted from the light curve. As an alternative result, these manual processes were mixed, i.e metadata was used in conjunction with the manually extracted features from the light curve.

With respect to the unbalanced techniques processes, the weights over the classes improved the results by ~ 0.1 on *F1 score* metric compared to undersampling technique. We suspect that is because if we delete some of the data we lose some patterns that can help the models to generalize better, by weighting we allow the model to learn for every data recollected as training set by setting the corresponding importance.

The results are shown by Table 3, which presents *F1 score* metric for the best representations of each technique, with 5 components for PCA and ICA, filling the missing data with zeros and assigning balanced weights to the classes (without subsampling the major class). Although the manual features extraction of the light curve (OwnFATS) has the advantage of do not require that the data sequences have the same length, since statistics are extracted independently, it turns out to be worse than the automatic techniques because they are adapted to the data (in the training process) and achieve to extract valuable information for the prediction.

After validate all the variants in the experimentation process and hyper-parameter selection the test results in Table 3 shows that the best *F1 score* was obtained using manual techniques, in which statistics and fixed light-curve features were extracted in addition to other manual computed features based on astronomy study (for the planet and the star). We obtained a performance of 88.3% for future classification, as *F1 score* metric indicated, meaning that probably 88.3% of the time the predictions will be completed and correct. The best trained model was Random Forest, where the Figure 2 presents the feature importance/relevance that this ensemble has on his process of predict the class. In the figure, it can be seen that the most relevant features belong to the object in study (object of interest). Particularly, the radius of the object is the one generate most impact on the prediction together with the period and the number of object in studies in that system. The features with less importance happens to be the ones extracted from the light curve, where the slope and second quartile are the ones with less impact.

The experimentation trying to make work the recurrent neural network was extensive but without success: we tested different representation on the input data, we varied the size of the window from 300 to 500 in which some features was extracted, we also varied the architecture of the network modifying the depth, changed the number of units, tried different optimizer (RMSprop and Adam), used different number of epochs as well as modified the batch size during the training. Unfortunately, no good results were achieved for any of the networks, being a little network with GRU the one with the best performance, 0.567 according to *F1 score*. Knowing the good performances of these models in other areas, we suspect that the sequence was too large and, for this reason, the statistics by window was not the best technique to summarize all the needed information for the proper learning of the neural network model.

The details of the precision and recall metrics on the classification over both classes in the test set can be found in the appendix. In Table 5 we report the classification on the *False Positive* class. It can be seen that the best model that can identify correctly this class based on this two metrics, i.e. the model that cover the most examples of that class and do this in a meticulous way (without including examples from other class), is the SVM RBF. This can be explained due to the RBF kernel, since it can fit boundaries on a flexible way and very tight to the data of that class. Similarly, in this Table it can be seen that the ICA

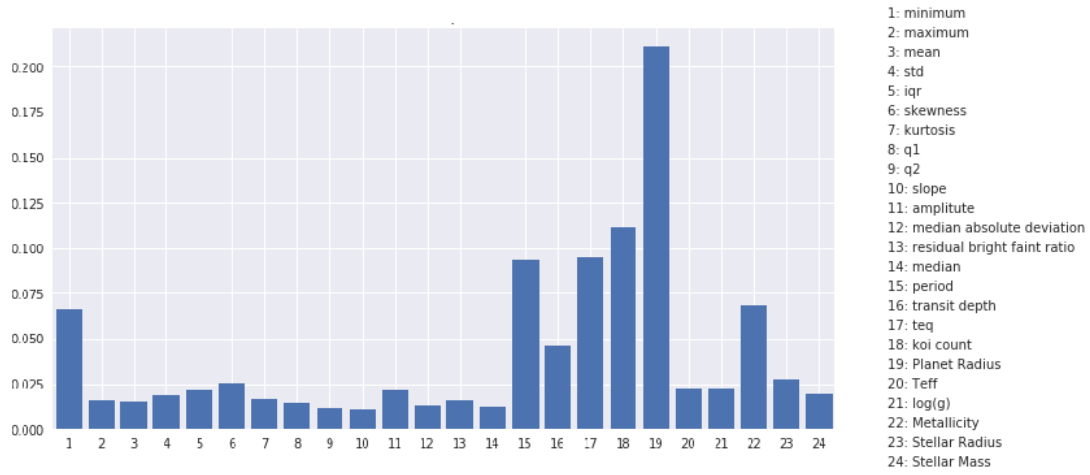


Figure 2: Feature importance of Random Forest model over the representation with the best performance (OwnFATS with metadata).

transformation applied over the frequency domain achieve the higher precision among all the techniques of features extraction, even compared to metadata, but with the trade-off of a small recall. This mean that it can only cover a small portion of the data of that class. In the same way, we can analyze the Table 6, i.e., the classification on the *Confirmed* class, showing lower scores that the other class, suggesting the difficulty on the exoplanet prediction problem. This could be because all of them do not have very similar features on the light curve or in the planet composition, making it difficult to detect and group all the samples of this class. However, the best model in the task of only detecting exoplanets (*Confirmed*) over the different representation of the data was Random Forest. This could be thanks to the flexibility (non-linearity) and low variance that achieve this ensemble over a single Tree and other learning models.

5.1 Final Results

After having performed all the corresponding tests and identified the best model on each label based on the metrics precision and recall, we show the predictions on the representation **OwnFATS + stellar & planet metadata** in Figure 3. The Random Forest model was selected for *Confirmed* class, with maximum depth 15 and the SVM RBF model with regularization parameter 100 for *False Positive* class. Therefore, we show the classification over the Kepler Object of Interest that are still being studied by the staff of NexScI on September of 2017, i.e. those object labeled as *Candidate*. Also, we report when the models disagree on assigning labels for the same object, labeling them as *Unclassified*, because the models cannot reach a consensus. This can be seen as an ensemble with two experts model on different classes, that gives a prediction when both models agree, representing that the expert model claim his class and the other support him. On the Table 7 (in the appendix) we show a sample of the labels predicted by the process mentioned, being these *Confirmed* or *False Positive* as appropriate. This table shows, for example that the system of the star Kepler 279, sheltering two confirmed exoplanet by NexScI (Kepler 279 b and Kepler 279 c), our models show that the third object in study **K01236.04** happens to be a valid exoplanet as the others orbiting the parent star. An opposite case is the system of the star Kepler 619, which also shelters two exoplanets (Kepler 619 b y Kepler 619 c): our models assign that the third object in study **K00601.02** result a false positive object. Also our models classify an entire system on study (**K01358.01**, **K01358.02**, **K01358.03** and **K01358.04**), tagging every object there as a valid exoplanet. A similar case can also be seen when it labels objects of interest on star with confirmed exoplanets, such as Kepler 763 b orbiting Kepler 763, where our models assigned two new brother exoplanets orbiting the same star yet the fourth object on study **K01082.02** is a false positive.

Finally we present a summary table on the assignment/labeling of the models in our work:

Total *Candidate*	**Learner**	**1791**
Subtotal *Confirmed*	Random Forest	975
Subtotal *False Positive*	SVM RBF	434
Unclassified		382

Table 4: Subtotal of candidate exoplanets by corresponding method.

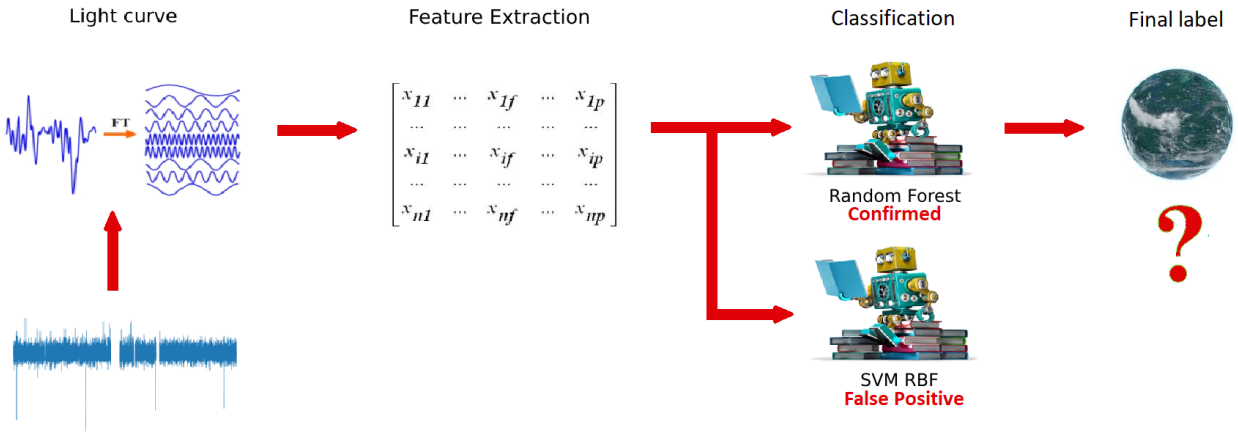


Figure 3: Labeling process for the Kepler dataset. The whitened light curves are transformed into the frequency domain (Fourier), in order to perform feature extraction based on the manual or automatic techniques presented on the work. Finally two selected learners (Random Forest and SVM RBF) classify the objects producing a final label when both agree.

6 Conclusion

We introduced a new method for refining the decision whether or not an object on study (KOI) is really an exoplanet using supervised automatic learning. By using different techniques focused on handling raw data (sequence of intensity of light) on machine learning and combining them properly, we reproduce the arduous and extensive work that experts perform when detecting or disconfirming an exoplanet on study. Based on the results of the feature engineering process, we indicate that the automatic techniques used to extract information from the light curve was not good enough compared to the metadata, which outperforms it respect to the score. This can be explained by an inadequate or naive choice of the methods for feature extraction, being too simple for the complex problem of coping with very diverse light curves in their morphology. Also the problem was complex regarding the execution time, due the computational cost of compute all the operations (feature extraction) over the very long sequence.

We also recognize that the different elements of the metadata used could have great impact on the results, because the choice was made only based on the description of the features informed by MAST. We expect to achieve better results in the near future by having the support of an expert in that area. Also as a future work, we plan to use the Mandel-Agol fit of a light curve because of its smoothness. Also, we will explore new methods for filling the missing values or some other techniques that can handle this properly, like Gaussian Processes. In the same line, we plan to use new and different techniques on the task of the feature extraction on the light curve.

Acknowledgment

This work was possible thanks to Chilean Virtual Observatory, ChiVO. We also thanks the Programa de Iniciación Científica PIIC-DGIP of the Federico Santa María University for funding this project.

References

[1] A. Wolszczan, "Confirmation of earth-mass planets orbiting the millisecond pulsar psr b1257+ 12," *Science*, vol. 264, no. 5158, pp. 538–542, 1994. [Online]. Available: http://dx.doi.org/10.1126/science.264.5158.538

[2] M. Solar, M. Araya, L. Arévalo, V. Parada, R. Contreras, and D. Mardones, "Chilean virtual observatory," in *Computing Conference (CLEI), 2015 Latin American*. IEEE, 2015, pp. 1–7. [Online]. Available: http://dx.doi.org/10.1109/CLEI.2015.7359465

[3] M. Mayor and D. Queloz, "A jupiter-mass companion to a solar-type star," pp. 355–359, 1995. [Online]. Available: http://dx.doi.org/10.1038/378355a0

[4] J. W. Richards, D. L. Starr, N. R. Butler, J. S. Bloom, J. M. Brewer, A. Crellin-Quick, J. Higgins, R. Kennedy, and M. Rischard, "On machine-learned classification of variable stars with sparse and noisy time-series data," *The Astrophysical Journal*, vol. 733, no. 1, p. 10, 2011. [Online]. Available: http://dx.doi.org/10.1088/0004-637X/733/1/10

[5] N. R. Lomb, "Least-squares frequency analysis of unequally spaced data," *Astrophysics and space science*, vol. 39, no. 2, pp. 447–462, 1976. [Online]. Available: http://dx.doi.org/10.1007/BF00648343

[6] C. Donalek, S. G. Djorgovski, A. A. Mahabal, M. J. Graham, A. J. Drake, T. J. Fuchs, M. J. Turmon, A. A. Kumar, N. S. Philip, M. T.-C. Yang *et al.*, "Feature selection strategies for classifying high dimensional astronomical data sets," in *Big Data, 2013 IEEE International Conference on*. IEEE, 2013, pp. 35–41. [Online]. Available: http://dx.doi.org/10.1109/BigData.2013.6691731

[7] I. Nun, K. Pichara, P. Protopapas, and D.-W. Kim, "Supervised detection of anomalous light curves in massive astronomical catalogs," *The Astrophysical Journal*, vol. 793, no. 1, p. 23, 2014. [Online]. Available: http://dx.doi.org/10.1088/0004-637X/793/1/23

[8] C. D. Dressing and D. Charbonneau, "The occurrence rate of small planets around small stars," *The Astrophysical Journal*, vol. 767, no. 1, p. 95, 2013. [Online]. Available: http://dx.doi.org/10.1088/0004-637X/767/1/95

[9] D. Dressing, Courtney D & Charbonneau, "The occurrence of potentially habitable planets orbiting m dwarfs estimated from the full kepler dataset and an empirical measurement of the detection sensitivity," *The Astrophysical Journal*, vol. 807, no. 1, p. 45, 2015. [Online]. Available: http://dx.doi.org/10.1088/0004-637X/807/1/45

[10] S. D. McCauliff, J. M. Jenkins, J. Catanzarite, C. J. Burke, J. L. Coughlin, J. D. Twicken, P. Tenenbaum, S. Seader, J. Li, and M. Cote, "Automatic classification of kepler planetary transit candidates," *The Astrophysical Journal*, vol. 806, no. 1, p. 6, 2015. [Online]. Available: http://dx.doi.org/10.1088/0004-637X/806/1/6

[11] S. E. Thompson, F. Mullally, J. Coughlin, J. L. Christiansen, C. E. Henze, M. R. Haas, and C. J. Burke, "A machine learning technique to identify transit shaped signals," *The Astrophysical Journal*, vol. 812, no. 1, p. 46, 2015. [Online]. Available: http://dx.doi.org/10.1088/0004-637X/812/1/46

[12] C. Mackenzie, K. Pichara, and P. Protopapas, "Clustering-based feature learning on variable stars," *The Astrophysical Journal*, vol. 820, no. 2, p. 138, 2016. [Online]. Available: http://dx.doi.org/10.3847/0004-637X/820/2/138

[13] A. Mahabal, K. Sheth, F. Gieseke, A. Pai, S. G. Djorgovski, A. J. Drake, and M. J. Graham, "Deep-learnt classification of light curves," in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2017, pp. 1–8. [Online]. Available: http://dx.doi.org/10.1109/SSCI.2017.8280984

[14] T. A. Hinners, K. Tat, and R. Thorp, "Machine learning techniques for stellar light curve classification," *The Astronomical Journal*, vol. 156, no. 1, p. 7, 2018. [Online]. Available: http://dx.doi.org/10.3847/1538-3881/aac16d

[15] R. Akeson, X. Chen, D. Ciardi, M. Crane, J. Good, M. Harbut, E. Jackson, S. Kane, A. Laity, S. Leifer *et al.*, "The nasa exoplanet archive: data and tools for exoplanet research," *Publications of the Astronomical Society of the Pacific*, vol. 125, no. 930, p. 989, 2013. [Online]. Available: http://dx.doi.org/10.1086/672273

[16] W. D. Pence, L. Chiappetti, C. G. Page, R. Shaw, and E. Stobie, "Definition of the flexible image transport system (fits), version 3.0," *Astronomy & Astrophysics*, vol. 524, p. A42, 2010. [Online]. Available: http://dx.doi.org/10.1051/0004-6361/201015362

[17] K. Mandel and E. Agol, "Analytic light curves for planetary transit searches," *The Astrophysical Journal Letters*, vol. 580, no. 2, p. L171, 2002. [Online]. Available: http://dx.doi.org/10.1086/345520

[18] M. Falk, F. Marohn, R. Michel, D. Hofmann, M. Macke, C. Spachmann, and S. Englert, "A first course on time series analysis : Examples with sas [version 2012.august.01]," 09 2012.

[19] S. Carpano, S. Aigrain, and F. Favata, "Detecting planetary transits in the presence of stellar variability-optimal filtering and the use of colour information," *Astronomy & Astrophysics*, vol. 401, no. 2, pp. 743–753, 2003.

[20] I. Nun, P. Protopapas, B. Sim, M. Zhu, R. Dave, N. Castro, and K. Pichara, "Fats: Feature analysis for time series," *arXiv preprint arXiv:1506.00010*, 2015.

[21] L. I. Kuncheva and W. J. Faithfull, "Pca feature extraction for change detection in multidimensional unlabeled data," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 1, pp. 69–80, 2014. [Online]. Available: http://dx.doi.org/10.1109/TNNLS.2013.2248094

[22] M. R. Gamit, K. Dhameliya, and N. S. Bhatt, "Classification techniques for speech recognition: a review," *International Journal of Emerging Technology and Advanced Engineering*, vol. 5, no. 2, pp. 58–63, 2015.

[23] M. Verleysen and D. François, "The curse of dimensionality in data mining and time series prediction." in *IWANN*, vol. 5. Springer, 2005, pp. 758–770. [Online]. Available: http://dx.doi.org/10.1007/11494669_93

[24] L. Cao, K. S. Chua, W. Chong, H. Lee, and Q. Gu, "A comparison of pca, kpca and ica for dimensionality reduction in support vector machine," *Neurocomputing*, vol. 55, no. 1, pp. 321–336, 2003. [Online]. Available: http://dx.doi.org/10.1016/S0925-2312(03)00433-8

[25] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4, pp. 411–430, 2000. [Online]. Available: http://dx.doi.org/10.1016/S0893-6080(00)00026-5

[26] K. Bae, S. Noh, and J. Kim, "Iris feature extraction using independent component analysis," in *International Conference on Audio-and Video-Based Biometric Person Authentication*. Springer, 2003, pp. 838–844. [Online]. Available: http://dx.doi.org/10.1007/3-540-44887-X_97

[27] F. J. Harris, "On the use of windows for harmonic analysis with the discrete fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, 1978. [Online]. Available: http://dx.doi.org/10.1109/PROC.1978.10837

[28] A. Grinsted, J. C. Moore, and S. Jevrejeva, "Application of the cross wavelet transform and wavelet coherence to geophysical time series," *Nonlinear processes in geophysics*, vol. 11, no. 5/6, pp. 561–566, 2004. [Online]. Available: http://dx.doi.org/10.5194/npg-11-561-2004

[29] B. Dasarathy, "Nearest neighbor norms: Nn pattern classification techniques," 1991.

[30] M.-L. Zhang and Z.-H. Zhou, "A k-nearest neighbor based algorithm for multi-label classification," in *2005 IEEE international conference on granular computing*, vol. 2. IEEE, 2005, pp. 718–721.

[31] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 215–242, 1958. [Online]. Available: http://dx.doi.org/10.1111/j.2517-6161.1958.tb00292.x

[32] V. N. Vapnik, *An overview of statistical learning theory*. IEEE, 1999, vol. 10, no. 5. [Online]. Available: http://dx.doi.org/10.1109/72.788640

[33] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1. IEEE, 1995, pp. 278–282.

[34] L. Fraiwan, K. Lweesy, N. Khasawneh, H. Wenz, and H. Dickhaus, "Automated sleep stage identification system based on time–frequency analysis of a single eeg channel and random forest classifier," *Computer methods and programs in biomedicine*, vol. 108, no. 1, pp. 10–19, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.cmpb.2011.11.005

[35] S. Hochreiter and J. Schmidhuber, "Lstm can solve hard long time lag problems," in *Advances in neural information processing systems*, 1997, pp. 473–479.

[36] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014, pp. 1724–1734. [Online]. Available: http://dx.doi.org/10.3115/v1/D14-1179

[37] C. Drummond, R. C. Holte *et al.*, "C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling," in *Workshop on learning from imbalanced datasets II*, vol. 11. Citeseer, 2003, pp. 1–8.

[38] G. King and L. Zeng, "Logistic regression in rare events data," *Political analysis*, vol. 9, no. 2, pp. 137–163, 2001. [Online]. Available: http://dx.doi.org/10.1093/oxfordjournals.pan.a004868

Appendix

	Learners			
	k-NN	*Logistic Regression*	*SVM RBF*	*Random Forest*
Fourier + PCA	P: 0.726 R: 0.809	P: 0.902 R: 0.283	P: 0.629 R: 1.000	**P: 0.789** **R: 0.736**
Fourier + ICA	P: 0.752 R: 0.728	P: 0.899 R: 0.285	**P: 0.933** **R: 0.332**	P: 0.788 R: 0.730
OwnFATS	P: 0.743 R: 0.695	P: 0.821 R: 0.441	**P: 0.890** **R: 0.395**	P: 0.827 R: 0.569
Planet metadata	P: 0.863 R: 0.857	P: 0.917 R: 0.830	**P: 0.927** **R: 0.817**	P: 0.914 R: 0.871
Stellar metadata	P: 0.781 R: 0.886	P: 0.806 R: 0.714	P: 0.818 R: 0.768	**P: 0.819** **R: 0.803**
OwnFATS + stellar & planet metadata	P: 0.860 R: 0.899	P: 0.919 R: 0.857	**P: 0.934** **R: 0.861**	P: 0.924 R: 0.884

Table 5: Scores of Precision (P) and Recall (R) on the *False Positive* class of learned models on the test set over the different representation on the data. In bold is the best model on each representation.

	Learners			
	k-NN	*Logistic Regression*	*SVM RBF*	*Random Forest*
Fourier + PCA	P: 0.597 R: 0.481	P: 0.438 R: 0.948	P: 0.000 R: 0.000	**P: 0.597** **R: 0.666**
Fourier + ICA	P: 0.562 R: 0.592	P: 0.438 R: 0.945	P: 0.458 R: 0.960	**P: 0.592** **R: 0.665**
OwnFATS	**P: 0.533** **R: 0.592**	P: 0.468 R: 0.836	P: 0.471 R: 0.917	P: 0.522 R: 0.798
Planet metadata	P: 0.763 R: 0.773	P: 0.755 R: 0.874	P: 0.745 R: 0.893	**P: 0.801** **R: 0.864**
Stellar metadata	**P: 0.755** **R: 0.585**	P: 0.600 R: 0.713	P: 0.649 R: 0.716	P: 0.681 R: 0.703
OwnFATS + stellar & planet metadata	P: 0.818 R: 0.756	P: 0.785 R: 0.874	P: 0.795 R: 0.898	**P: 0.820** **R: 0.879**

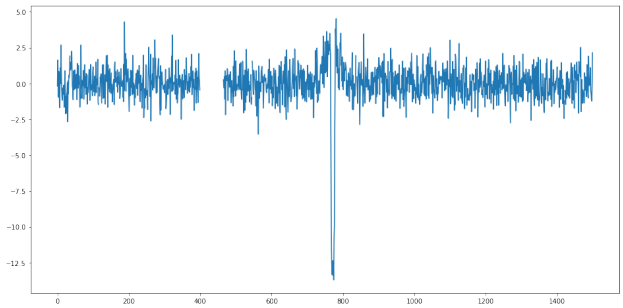
Table 6: Scores Precision (P) and Recall (R) to the *Confirmed* class of learned models on the test set over the different representation on the data. In bold is the best model on each representation

KOI name	**Disposition**	**Confirmed on that system**	**Star**
K00601.02	*FALSE POSITIVE*	2/3	Kepler 619
K00750.02	*UNCLASSIFIED*	1/3	Kepler 662
K01082.01 K01082.02 K01082.04	*CONFIRMED* *FALSE POSITIVE* *CONFIRMED*	1/4	Kepler 763
K01236.04	*CONFIRMED*	2/3	Kepler 279
K01358.01 K01358.02 K01358.03 K01358.04	*CONFIRMED* *CONFIRMED* *CONFIRMED* *CONFIRMED*	0/4	-
K01750.02	*CONFIRMED*	1/2	Kepler 948
K02064.01	*UNCLASSIFIED*	0/1	-
K02420.02	*CONFIRMED*	1/2	Kepler 1231
K02578.01	*FALSE POSITIVE*	0/1	-
K02828.02	*FALSE POSITIVE*	1/2	Kepler 1259
K03444.03	*UNCLASSIFIED*	0/4	-
K03451.01	*UNCLASSIFIED*	0/1	-
K04591.01	*FALSE POSITIVE*	0/1	-
K05353.01	*FALSE POSITIVE*E	0/1	-
K06267.01	*CONFIRMED*	0/1	-
K06983.01	*CONFIRMED*	0/1	-
K07279.01	*CONFIRMED*	0/1	-
K07378.01 K07378.02	*CONFIRMED* *CONFIRMED*	0/2	-
K07434.01	*FALSE POSITIVE*	0/1	-
K08082.01	*CONFIRMED*	0/1	-

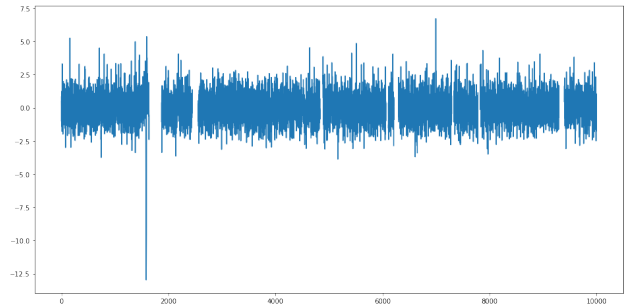
Table 7: This table show some of the predictions assigned as much by **OwnFATS + stellar & planet metadata** with Random Forest classifier for the task of exoplanet detection (*Confirmed*), as by **OwnFATS + stellar & planet metadata** with SVM RBF for the task of non-exoplanet detection (*False Positive*). It should be mentioned that the column **Confirmed on that system** count the ammount of confirmed exoplanets on the date of the study (September 2017), as long as **Star** is the name of the Parent star in the system; The ones with no information doesn't show this value. More can be found on https://github.com/FMena14/ExoplanetDetection

Samples of the input data

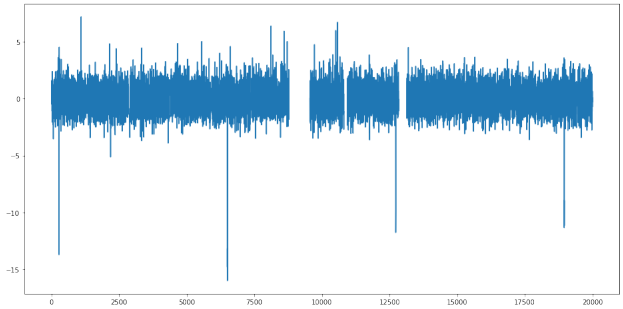
Light curve in raw data



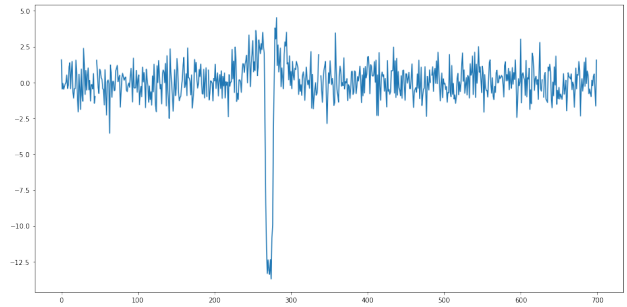
Light curve 1



Light curve 2

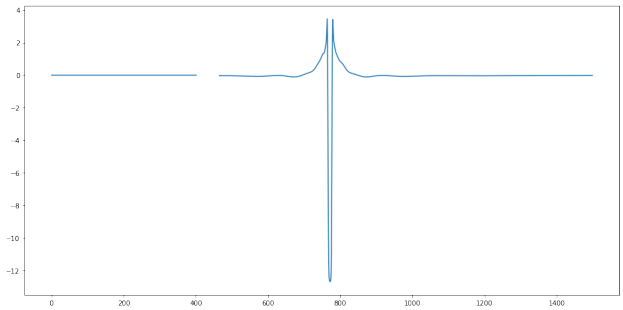


Light curve 3

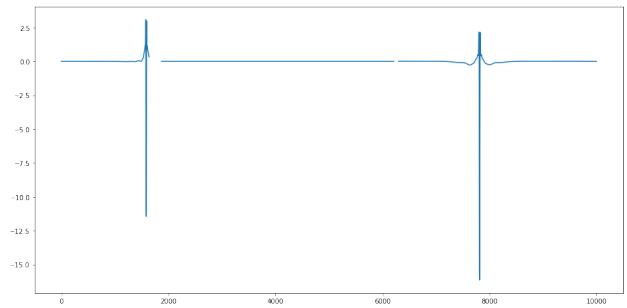


Light curve 4

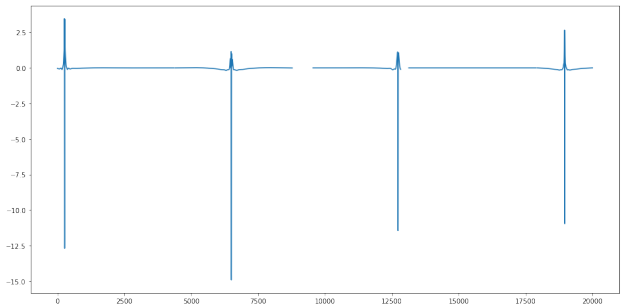
Mandel-Agol light curve fit



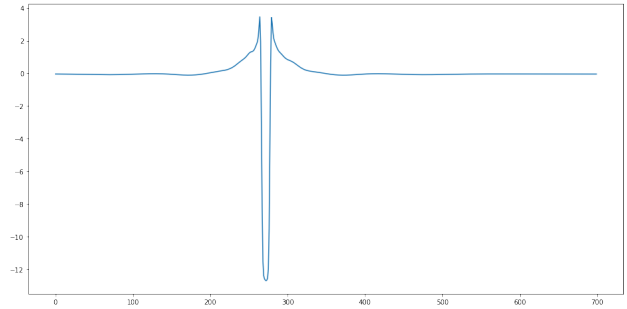
Model 1



Model 2



Model 3



Model 4