Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

A Novel Approach for Exoplanet Classification on Kepler Light Flux Data

[1]Abhishek Jha*, [2]Aayush Bajaj, [3]Lakshay Vashisth, [4]Dr. V. K. Saini
Department of Information Technology, Maharaja Agrasen Institute of Technology, New Delhi, India
[1]jhabhishek3797@gmail.com,[2] dev.aayushbajaj@gmail.com, [3]lakshaylv97@gmail.com, [4] vinaysaini@mait.ac.in

Article Info
Page Number: 1128-1134
Publication Issue:
Vol. 71 No. 3s (2022)

Article History
Article Received: 22 April 2022
Revised: 10 May 2022
Accepted: 15 June 2022
Publication: 19 July 2022

Abstract— NASA's Kepler mission to search for extrasolar planets has collected data from hundreds of thousands of star systems and has discovered nearly 1000 confirmed exoplanets in addition to over 3000 unconfirmed candidates. The mission discovers exoplanets using transit photometry, which detects the transit of a planet in front of a star as transient drops in stellar intensity. We propose a novel machine learning based approach to detect these exoplanets from Kepler's light flux dataset using robust preprocessing methods. We observed the dataset to be a highly imbalanced one having positive classes totalling less than 1% of the entire dataset and further, there was an inherent non-uniformity in the recorded data for positive samples. Our novel approach comprises applying Normalisation followed by Gaussian Smoothing and Fourier Transformation on this dataset before classifying it using Linear Support Vector Classifier. On such challenging data, our approach pushes the test F1 score to 1.0. While pursuing this objective of maximising test precision and recall, we also perform a comparative study of several data transformations in combination with various state-of-the-art modelling algorithms. We also compare our results with previous work and highlight the improvements obtained via our methodology.
Index Terms—Fourier transform, Imbalance classification, Kepler light flux dataset, Linear Support Vector Classifier

I.INTRODUCTION

NASA's Kepler spacecraft spent over four years collecting data on hundreds of thousands of star systems in search of exoplanets. This data was collected by taking images of a constant patch of space every half hour on a continuous basis for 3 months. Every 3 months the spacecraft was recalibrated and this cycle was repeated for 17 quarters before the Kepler spacecraft failed, ending its primary mission. From these images, the pixels corresponding to stars were identified, and the pixel location and intensity values were recorded over time. Taken together, this generated a series of light curves for each tracked star, from which exoplanets could be detected using transit photometry by looking for characteristic intensity dips as planets transit in front of the star [1][2].

There are other methods as well for hunting exoplanets, such as the radical-velocity-based method, in which rather than looking for signs of planets, the concerned star is observed for movements and its velocity is used for determining the presence of a planetary system around it [3]. However, this method and others lie beyond this paper's scope as our main focus is to find an elegant transit-based method which can learn from the aforementioned "dips".

No matter which method you may choose this is a very challenging problem as out in space, there are a lot more non-exoplanet stars than there are with orbiting exoplanets and this imbalance gets reflected in our dataset as well. As a result, our methodology must be extremely

accurate in identifying the True Positives (exoplanets) as well as in discarding the True Negatives (non-exoplanets). The entirety of the analysis, experiments and modelling has been carried out keeping this goal in mind which is to maximise precision as well as recall. To achieve the stated objective the data has been subjected to intense preprocessing and feature engineering to create a distinction between the two classes. Thereafter, a number of algorithms have been trained on this refined dataset and evaluated on the basis of the confusion matrix.
The paper is organised into 5 sections. Section 2 discusses our methodology, the dataset, the transformations applied and the model training. Section 3 contains the final results as well as a comparative evaluation of implemented approaches. In section 4 we collate the results of our approach with previously attempted works. Finally, we conclude and discuss the future steps in Section 5.

II.METHODOLOGY

Machine Learning methods are extensively used in scientific areas to build classifiers. In our case, we are building a binary classifier which will separate each time-series photometry, into classes “Exoplanet” and “Non-Exoplanet”. As opposed to deep learning techniques applied by [5] for planet detection where features were determined automatically, we’ve derived features from smart transformations and these are provided as inputs to our array of models for a comparative study.
We use 5 different models Artificial Neural Network, Decision Trees, Random Forest, LinearSVC, XGBoost on
the Kepler dataset to establish a baseline and do a comparative study on them. Each training stage comprised –

1. Processing and labelling raw data
2. Extracting features by Transformations
3. Model Training

The workflow is shown in fig. 1. More details on preprocessing and training steps can be found in the subsequent sections.

A.Preparing and labelling the Kepler dataset

The data originates from the observations made by NASA Kepler Space Telescope. From an operational perspective, the Kepler telescope used to conduct campaigns during which it used to focus on a certain area of space and observed any fluctuations in the light emitting from that region for around 80 days translating to around 3197 Light Flux values. Being extremely photosensitive, it was able to record such disturbances quite easily. The recorded data was then beamed down to earth and after a little cleaning, it was open-sourced via the Mikulski Archive [6].

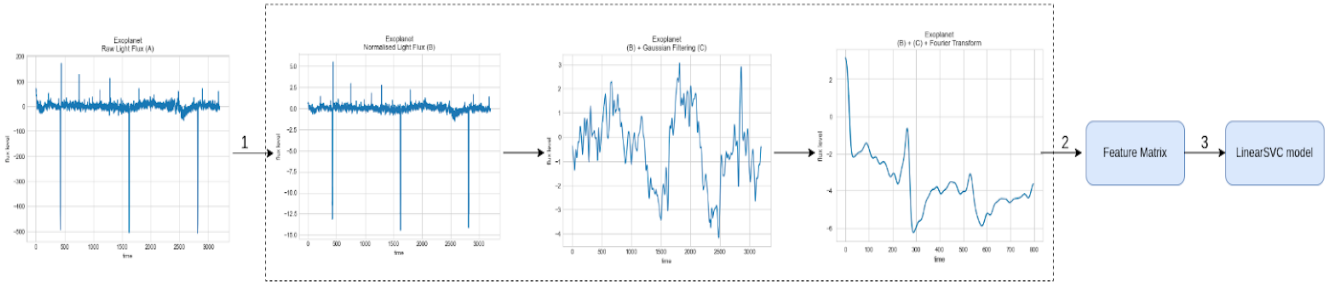

Fig. 1
In the first step (1) the raw light curves are preprocessed using a range of transformations including Normalization, Gaussian Smoothing and Fourier Transformation in a sequential manner. In the second step (2) a feature matrix is constructed which contains processed signals from each datum. Consequently, in the third and final step (3) the feature matrix is used as the input for a classification algorithm and the model is trained.

So effectively, the data describes the change in light intensity of thousands of stars. Observations have been labelled as either 1 or 2, 2 indicating the presence of at least one exoplanet while 1 indicating none. The data has been split into training and test set, with the former containing 5087 observations out of which 37 are confirmed exoplanet stars and the latter having 570 observations with only 5 positive samples, depicting the imbalance discussed earlier [7].

B.The Transformations

To make the dataset apt for modelling, a series of transformations were applied to the dataset in order to get distinguishing features for each of the classes.

1)Normalisation

Also known as spatial sign preprocessing, it scales input vectors individually to unit norm as seen in fig. 2. For L2, it is given by:

$$||x||_2 = \sqrt{\sum_i^n x_i^2} \qquad (1)$$

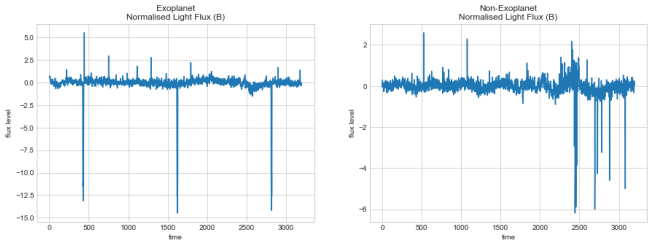

Fig. 2
Except the range of y-axis, not much changed in terms of pattern of light flux values of the two classes.

2)Gaussian Filtering

Here, Gaussian filtering refers to convolving over our time-series data with a Kernel, which is of the shape of a Gaussian curve. Gaussian Kernel is given by this equation:

$$K(x^*, x_i) = \exp\left(\frac{-(x^* - x_i)^2}{2b^2}\right) \qquad (2)$$

where x* is a list of datapoints, xi is the datapoint for which we are calculating the filtered value and b defines the width of kernel.

Post normalisation, we proceed for all the datapoints one by one, generating a new value that is a function of the original value of that datapoint and the values of surrounding datapoints. Once this process is repeated for the entire dataset we get a smoothened curve with increased signal to noise ratio as seen in fig. 3. In a time-series data this helps us better see patterns and trends.

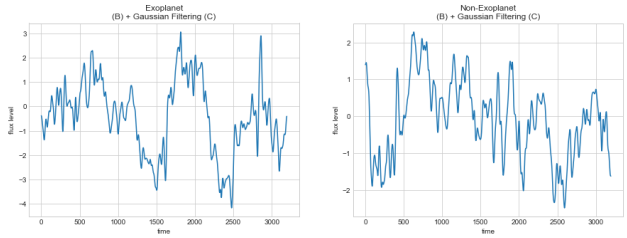

Fig. 3
After two rounds of transformation, the light curves still looked similar for the two classes.

3)Fourier Transform

The main application of Fourier Transform is to convert a time-domain function f(t) into its corresponding frequency-domain representation $F(\omega)$. t in original function denotes time while in ω transformed function denotes frequency. It is given by the following equation:

$$F(\omega) = \int_{-\infty}^{+\infty} f(t) e^{-2\pi i \omega t} dt \qquad (3)$$

Just like Gaussian Smoothing, It is yet another useful tool for de-noising a signal and finding the harmonics of waveform. It is specially useful if the data is periodic and there is some meaningful information encoded in this periodicity. Considering our data is supposed to exhibit a similar pattern in the light flux signals of exoplanets, decomposing the data into sinusoidal waveforms after applying the last two operations helped us get discernible features as seen in fig. 4.

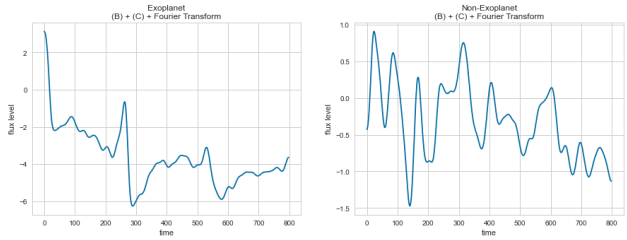

Fig. 4
After Fourier Transformation on signal of length N, we get back a symmetrical signal of length N/2. The distinction between the two classes becomes quite clear after this step.

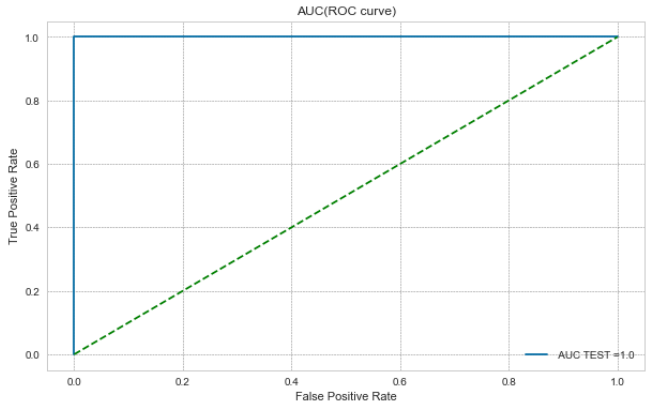

Fig. 5

III.RESULTS

We trained the algorithms discussed in the last section on each of the transformed sets of data, following which they were evaluated on the test data of 570 samples. The best results were produced with Linear Support Vector classifier on the data obtained after the application of the three transformations, as discussed earlier. The results are presented in fig. 5.

While the Linear SVC model was able to utilise the transformed features to their full extent, this was not true for the rest of the algorithms. Test F1 scores of all the combinations have been summarised in Table 1.

IV.PREVIOUS WORK COMPARISON

Table 2 exhibits the improvement achieved by our method over previous works [8][9]. While the other works stressed more over the choice of the models, the focus of our approach was skewed in the favour of getting more meaningful features out of the data. With this motivation, we were able to better the results and achieve perfect test Precision and Recall for positive samples.

Table I
Test F1 Score - Comparative Study

Models	Raw Data (A)	Normalised Data (B)	(B) + Gaussian Smoothing (C)	(B) + (C) + Fourier Transformation
ANN	0	0	0	0
Decision Tree	0.036	0.5	0	0
Linear SVC	0	0.11	0	1
Random Forest	0	0	0.12	0
XGBoost	0	0	0	0

Table II

Test scores for exoplanet class

	Models	Precision	Recall
Dreborg, Linderholm, Tiensuu and Örn, 2019	CNN SVM	0.769 0.571	1 0.8
Singh and Kumbhare, 2022	CNN SVM	0.03 0.01	0.4 0.4
Our method	Linear SVC	1	1

CONCLUSION

With the advent of Deep Learning models, it has increasingly been the trend to view problem statements from the perspective of modelling algorithms and expect these models to learn the intricate patterns in the data by themselves, while completely ignoring the importance of pre-processing in such exercises. In this paper, we were able to find perfect results and score improvement over previous works with a conventional Machine Learning algorithm, Linear SVC by simply subjecting the data to rigorous analysis and feature engineering. With discernible features obtained from our transformed data, Linear SVC was able to distinguish between the two classes with little effort, thus proving the potential of such elegant methods. Moreover, being a lightweight model, our approach doesn't require the availability of special hardware like GPUs and the training takes even less than 30 seconds on a quad-core CPU system [4].
While these results are extremely encouraging, solutions developed with limited data may commit mistakes on unseen data. Hence, such methods must be used alongside human supervision until they are robust enough. Nonetheless, the core idea of our approach is completely reliable and can be scaled to big systems and more complex problem statements to get desirable results and reduce manual effort

REFERENCES

1. "Exploring Exoplanets with Kepler Activity." NASA/JPL Edu. Accessed April 10, 2019. https://www.jpl.nasa.gov/edu/teach/activity/exploring-exoplanetswith-kepler/.
2. Jin, X. and Glass, D., 2014. Searching for exoplanets in the Kepler public data. Cs229.stanford.edu.
3. "Exoplanet Exploration: Planets Beyond Our Solar System. Accessed April 11, 2019. https://exoplanets.nasa.gov/.
4. Malik, A., Moster, B. and Obermeier, C., 2021. Exoplanet detection using machine learning. Monthly Notices of the Royal Astronomical Society.
5. Alaria, S. K., A. . Raj, V. Sharma, and V. Kumar. "Simulation and Analysis of Hand Gesture Recognition for Indian Sign Language Using CNN". International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 4, Apr. 2022, pp. 10-14, doi:10.17762/ijritcc.v10i4.5556.
6. Bulla, P. . "Traffic Sign Detection and Recognition Based on Convolutional Neural Network". International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 4, Apr. 2022, pp. 43-53, doi:10.17762/ijritcc.v10i4.5533.

7. Shallue, C. and 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(2), p.94."K2." STScI. Accessed April 10, 2019. http://archive1.stsci.edu/k2.
8. "Exoplanet Hunting in Deep Space." Accessed April 10, 2019. https://kaggle.com/keplersmachines/kepler-labelled-time-series-data.
9. Dreborg, S., Linderholm, M., Tiensuu, J. and Örn, F., 2019. Detecting exoplanets with machine learning. Uppsala University.
10. Singh, A. and Kumbhare, V., 2022. Detection of Exoplanets using Machine Learning. International Journal of Research Publication and Reviews, pp.1007-1015.
11. Malla, S., M. J. . Meena, O. . Reddy. R, V. . Mahalakshmi, and A. . Balobaid. "A Study on Fish Classification Techniques Using Convolutional Neural Networks on Highly Challenged Underwater Images". International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 4, Apr. 2022, pp. 01-09, doi:10.17762/ijritcc.v10i4.5524.
12. Kadhim, R. R., and M. Y. Kamil. "Evaluation of Machine Learning Models for Breast Cancer Diagnosis Via Histogram of Oriented Gradients Method and Histopathology Images". International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 4, Apr. 2022, pp. 36-42, doi:10.17762/ijritcc.v10i4.5532.