

A machine learning approach to exoplanet prediction through pattern analysis using the transit method

Amy Zheng

Introduction

In recent years, there has been an influx in discovery in the field of exoplanet detection, especially with the large amounts of data being gathered by NASA missions, such as the Kepler mission or the K2 mission. Currently, there have been thousands of exoplanets discovered using a variety of different methods, the most common ones being:

- Direct imaging: infrared wavelengths emitted by the planets are observed and recorded [1].
- Radial velocity: slight movement by the star in response to its orbiting planets is recorded [2].
- Transit photometry: the light flux in a star recorded as a result of a planet transiting the star. The graph of the system displays a notable dip in the data when light from the star is blocked [3].

Though each method has made valuable impacts in the field, the transit method demonstrates the most efficiency, with 2951 exoplanet candidates discovered using this method, making it a viable test method.

However, despite the revolutionary progress made towards discovering new exoplanets, much of the data gathered by the Kepler missions have yet to be analyzed and classified, especially the data for exoplanet habitability. Although machine learning has been used for the purpose of exoplanet transit detection in previous studies, such as Hippke and Heller's 2019 paper on the transit least-squares algorithm, habitability is not often assessed through these methods [4]. With manual observations being a tedious and time-consuming task, this project seeks to develop a classical machine learning method that is both accurate and efficient in order to analyze the transit data both for candidates and habitability.

Methods

The data gathered from the Kepler Mission launched by NASA, specifically, the “Kepler Objects of Interest” (KOI) dataset, was obtained from the NASA Exoplanet Archive for the model. The archive currently contains the data for 9564 entries, each entry containing information on the targets’ properties and stellar parameters. Each entry is also labeled as follows:

- “Confirmed”, when a candidate has been confirmed to be an exoplanet
- “False positive”, when a candidate had similar features identified as an exoplanet, but ultimately a deeper analysis demonstrates it is not an exoplanet
- “Candidate”, when a candidate’s status is yet to be determined

Of the 9564 entries, I downloaded the records for 8197 entries, as there were occurrences where a single planet, associated with a singular Kepler ID, would have had data for multiple light curves recorded. Additionally, entries were also eradicated if they did not contain data for the factors I sought to test. In total, I obtained the data of 1934 confirmed candidates, 4452 false positive candidates, and 1811 candidates that have yet to be determined.

Each record was downloaded along with the MAST (Mikulski Archive for Space Telescopes) data from the raw light curve. The light curve is considered time series data, and because the model I seek to implement is unable to make predictions based on time series data efficiently, feature extraction was performed using FATS (Feature Analysis for Time Series) to obtain the necessary orbital parameters for the training process.

The Python-based package, *Lightcurve*, was also initiated on the raw lightcurve to further process the data through folding the curve and removing distracting variability and noise (Fig. 2). By folding the curve, the light curve graph displays the flux in relation to the time parameter represented by period (epoch), instead of time in days. This allows for a more accurate simulation of the movement of the planet when it crosses in front of a star. The following features that are significant towards the training process were then acquired: orbital period, transit epoch, transit depth, equilibrium temperature, insolation flux, number of planets, stellar effective temperature, surface gravity, metallicity, radius, and mass.

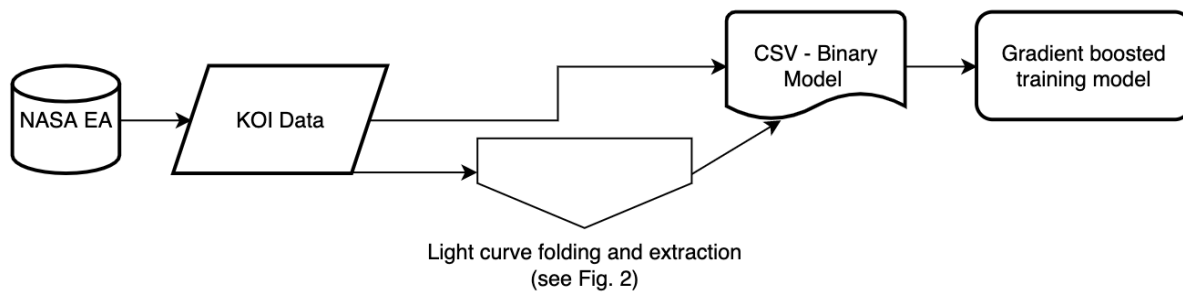


Fig. 1. Overview of procedure used to obtain data, extract features, and train model.

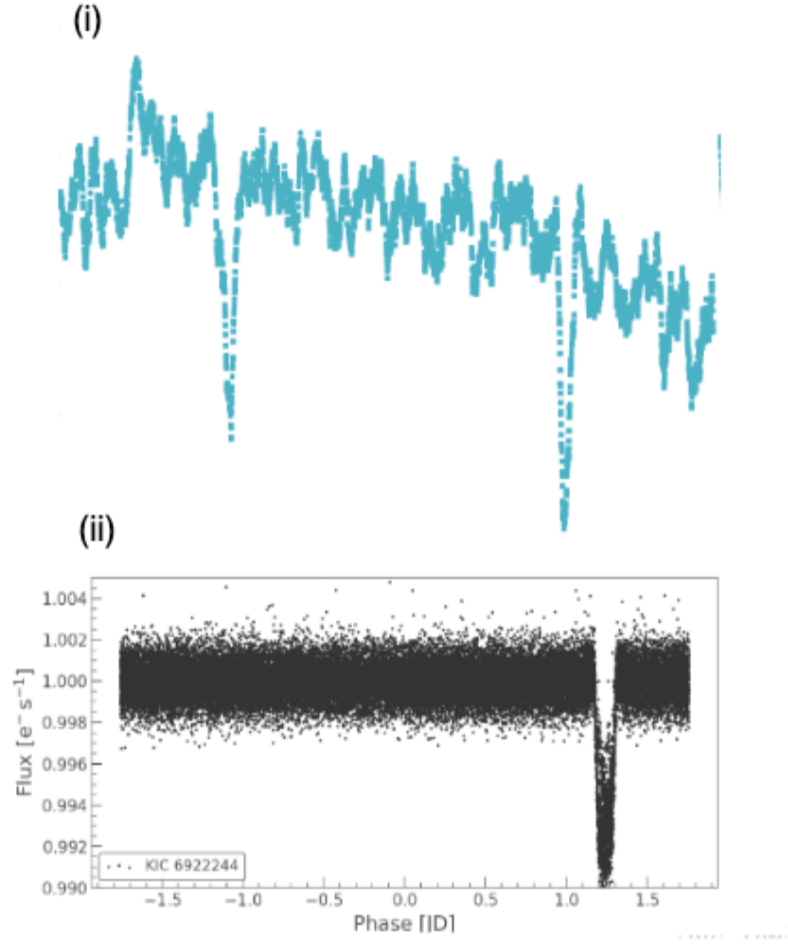


Fig. 2. (i) The raw light curve provided by MAST vs. (ii) the processed one used in FATS.

To train the model, the gradient boosted tree, *lightGBM*, was used. Because it is a binary classifier, I labeled the data obtained from the KOI dataset, with class 1 denoting a confirmed candidate and class 0 being a false positive. The analyzed data of 1934 confirmed candidates and 4452 false positive candidates were divided into the split of 60% train, 20% validation, and 20% test. The hyperparameters that were necessary to tune for this model included the number of leaves and the maximum depth of the tree in order to prevent both under and overfitting.

Following the training of the model, the test dataset, containing the 1811 planet candidates that have an undetermined status, was tested using the model to provide predictions for the status of each planet candidate.

The second stage of the model was to test each exoplanet for habitability. The data for a total of 59 habitable exoplanets, including their mass, radius, flux, equilibrium temperature, and orbital period, were downloaded through the “Habitable Exoplanets Catalog” [5]. This database contains exoplanet data from the NASA Exoplanet Archive of planets within the stellar habitable

zone. Among the 59 habitable exoplanets, 23 came from the KOI Archive. The 59 habitable exoplanets and the 1911 inhabitable confirmed candidates were then split for train, validation, and testing.

The *lightGBM* framework was then implemented again with the second round of data. The test set, which included the exoplanet candidates that had a score of 80% and above to be a confirmed exoplanet was used for the test dataset. In order to measure the model performance, results for accuracy and area under the receiver operating characteristic curve (AUC) were obtained.

Results

Upon tuning the hyperparameters, using a maximum depth of 15, the accuracy measure for the model consistently remained around 88%, successfully exhibiting the efficiency of the machine learning methods used for exoplanet detection (Fig. 3). Additionally, the validation set yields an AUC score of 95%.

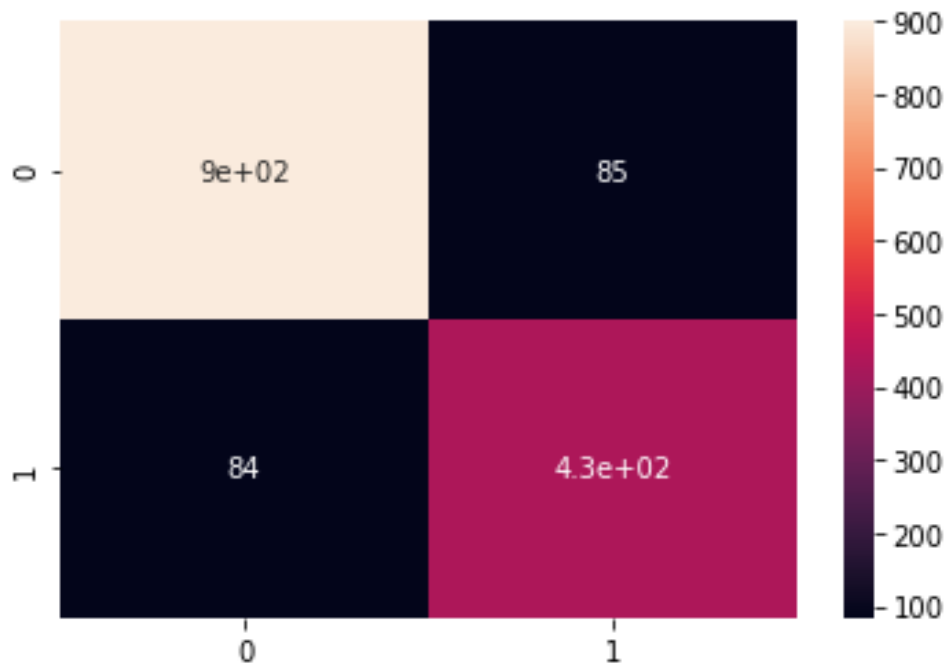


Fig. 3. The confusion matrix displaying the accuracy of the model

From the model, the planet that scored the highest was Kepler ID 8804283 which scored 99.11%, indicating a high probability of the candidate being a valid exoplanet. However, Kepler ID 7376983, which scored 97.01%, could possibly have an even higher probability of being a valid exoplanet because of its correlation with the other candidates in the same system (Fig. 4). The 3 other objects in the system scored 94.94%, 93.17%, and 88.56% respectively.

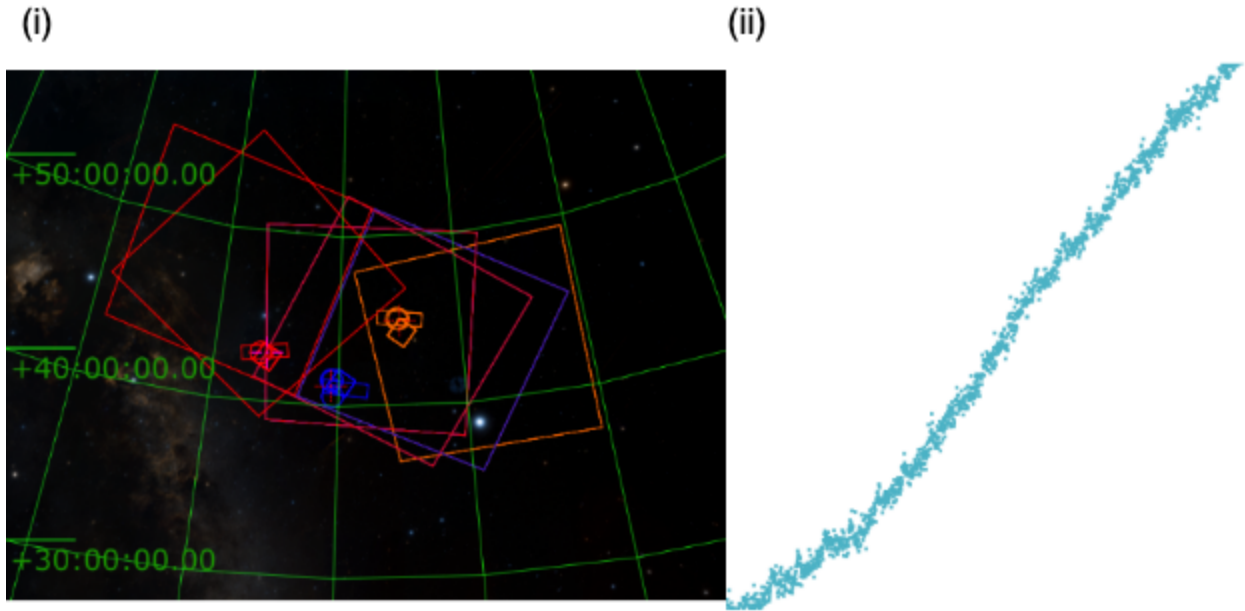


Fig. 4. (i) The location of Kepler ID 7376983 and system. (ii) The light curve of Kepler ID 7376983 provided by MAST

In the test for habitability, Kepler ID 11462341 yielded the highest score with an 81.15%. It is a planet candidate that has very similar features in comparison to Earth, particularly in radius and flux. It has a planetary radius that is 1.45 times the size of Earth, as well as an insolation flux of 2.21 (Fig. 5).

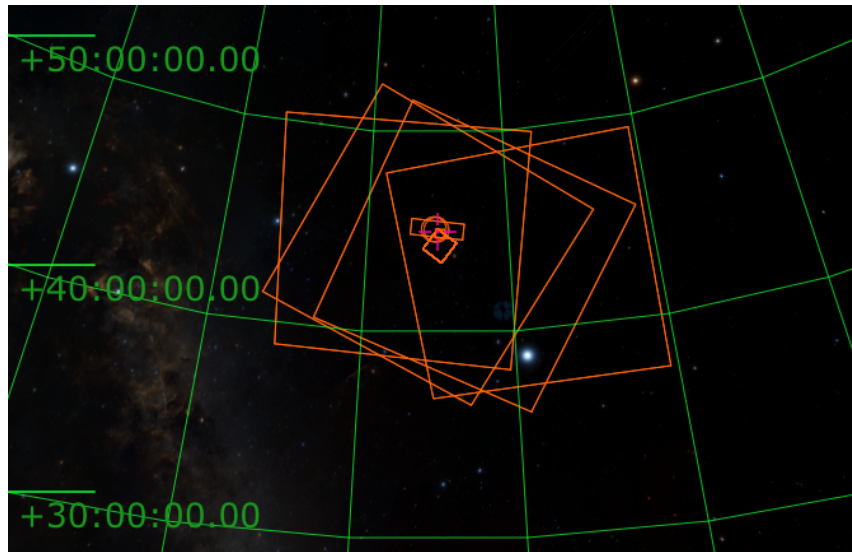


Fig. 5. The location of Kepler ID 11462341 provided by MAST

Discussion

The machine learning model used demonstrates the efficiency of a predictive model when detecting exoplanets through transit light curves. The accuracy and AUC scores indicate valid performances through the model. However, because the KOI dataset remains imbalanced, accuracy is not ideal for the purpose of evaluating the results [6]. Especially when testing for habitability, given the scarcity of possibly habitable exoplanets (around 3.1% of the total habitability dataset), the accuracy measure may not be a complete measure of the model's performance. In contrast, AUC considers the possibilities across all thresholds of data, allowing for a more comprehensive evaluation of the model.

There are also limitations to the Kepler data given the uncertainty and imbalances within the database. As a result, the hyperparameters were tuned to focus on receiving a higher AUC score [3]. In Scannell's 2021 study, they used Bayesian model-based optimization to optimize hyperparameters through the process of training [7]. Instead of using the grids search algorithm to test "poor-performing choices", the implementation of this method into my process could greatly assist in the tuning of hyperparameters.

Conclusion

Through the results obtained from the model, it can be concluded that machine learning methods can serve as a viable source for exoplanet detection.

There remains a level of uncertainty when it comes to machine learning models used in the classification of exoplanets. It is unclear how a certain model may perform when presented with new and unfamiliar data. In this aspect, manually vetting the data proves superior. However, this process is very tedious, especially when it comes to large amounts of data that are being gathered through the missions. The classical machine learning approach provides an accessible yet efficient method towards the deeper analysis of exoplanets, yielding straightforward and easily interpreted results. It could potentially aid the vetting experts in the study of these exoplanets, leading to accelerated discoveries in the field. Especially when more data is gathered by the missions and the model can be trained with more data, the model could be repeatedly trained and tested to obtain increasingly accurate results.

In terms of the exoplanet habitability training and testing, each exoplanet's proximity to M dwarfs could be considered as most exoplanets in the habitable zone can be found relatively close in distance to the M dwarf as shown in Schweiterman's study in 2019 [8]. Although the habitable zone from the host star was taken into consideration, more external factors, such as the presence of liquid water and carbon dioxide concentrations could be further examined in the future.

References

- [1] Chauvin, G. 4 October 2018. Direct Imaging of Exoplanets at the Era of Extremely Large Telescopes. arXiv:1810.02031. 12 December 2021.
- [2] Fischer, D., et.al. January 2011. Radial Velocity Techniques for Exoplanets. 12 December 2021
- [3] Malik, A., et.al. 5 March 2021. Exoplanet Detection using Machine Learning. <https://arxiv.org/pdf/2011.14135.pdf>. 21 October 2021.
- [4] Hippke, M., et.al. 22 January 2019. Optimized transit detection algorithm to search for periodic transits of small planets. <https://www.aanda.org/articles/aa/pdf/2019/03/aa34672-18.pdf>. 22 November 2021.
- [5] Kane, S., et.al. 2 September 2016. A Catalog of Kepler Habitable Zone Exoplanet Candidates. arXiv:1608.00620. 18 November 2021.
- [6] Carrington A., et.al. 21 March 2021. Deep ROC Analysis and AUC as Balanced Average Accuracy to Improve Model Selection, Understanding, and Interpretation. <https://arxiv.org/pdf/2103.11357.pdf>. 28 November 2021.
- [7] Scannell, Natasha. May 2021. The Search for Life: Exoplanet Detection with Deep Learning. <https://dc.uwm.edu/cgi/viewcontent.cgi?article=3730&context=etd>. 03 December 2021.
- [8] Schwieterman, E., et.al. 10 June 2019. <https://doi.org/10.3847/1538-4357/ab1d52>. 03 December 2021.