Searching for exoplanets using machine learning

Martin Mohan
x18191339
MSc/PGDip in Data Analytics
x18191339@student.ncirl.ie
20 October 2019

Abstract

This report reveals a need in the general astronomical community to improve machine learning methods and examines machine learning in exoplanet discovery in more detail. Machine learning is increasingly being used in the search for new planets and was used extensively on the Kepler mission which was searching for earth like planets. On the Kepler mission NASA used a random forest method to help predict possible planetary candidates. This was very important as Kepler produced an overwhelming amount of data. The accurate identification of potential candidates meant time was not spent needlessly trying to confirm non-existent planets with overstretched resources. The machine learning algorithm would vet out events such as instrument noise and binary stars.
After NASA had made public the Kepler data an engineer from Google's brain team examined the data using a different machine learning method called Convolutional Neural Networks (CNN) and found two as yet undiscovered planets. Only 670 of the 200,000 planets were tested using this CNN method which suggests there may still be more planets hidden in the Kepler data. This report suggests following up on this research and testing other machine learning algorithms on the data.

Index Terms

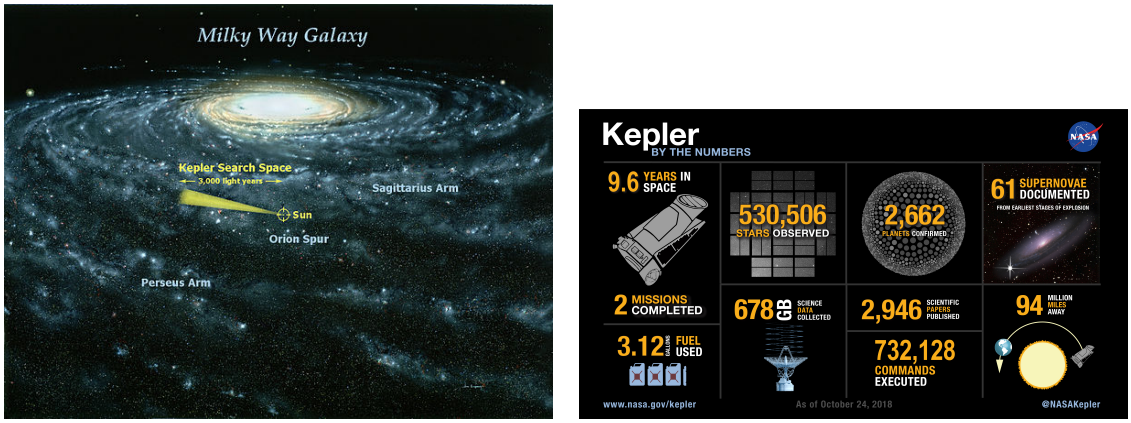
Extra solar planet, Exoplanet, Kepler, NASA, machine learning

CONTENTS

I. Introduction

The amount of data being gathered in astronomy is increasing every year and threatens to overwhelm the astronomers who try to process it. This wealth of data will only increase in years to come and astronomers are increasingly looking towards machine learning for help. This paper investigates machine learning methods for handling astronomical data using data from the Kepler mission.

The primary goal of the Kepler mission was to search for Earth-size planets in the habitable zones of solar like stars .The Kepler spacecraft stared at a 100 sq. degree patch of sky near Cygnus see Fig 1 (a) in order to measure the brightness variations of just over 200,000 stars. During its over nine and a half years of service, Kepler observed 530,506 stars and detected 2,662 planets see Fig 1 (b). Kepler showed there are more planets than stars in our galaxies.



(a) (b)

Fig. 1. Kepler Mission

Kepler used the Transit method see Fig 2 (a) to discover planets but usually another method is needed to confirm the planet such as Radial Velocity method see Fig 2 (b). The exoplanet discovery methods are listed below...

A. Exoplanet Discovery Methods

1) Transit Method: 2 (a) Transit photometry is currently the most effective and sensitive method for detecting extrasolar planets, particularly from an observatory in space. It accounted for 76.88% of all planet discoveries. Transit method involves detecting a slight dimming of a signal when a planet passes in front of a sun [1].Transiting gives astronomers a good estimate of the orbiting planet's size, but not its mass.

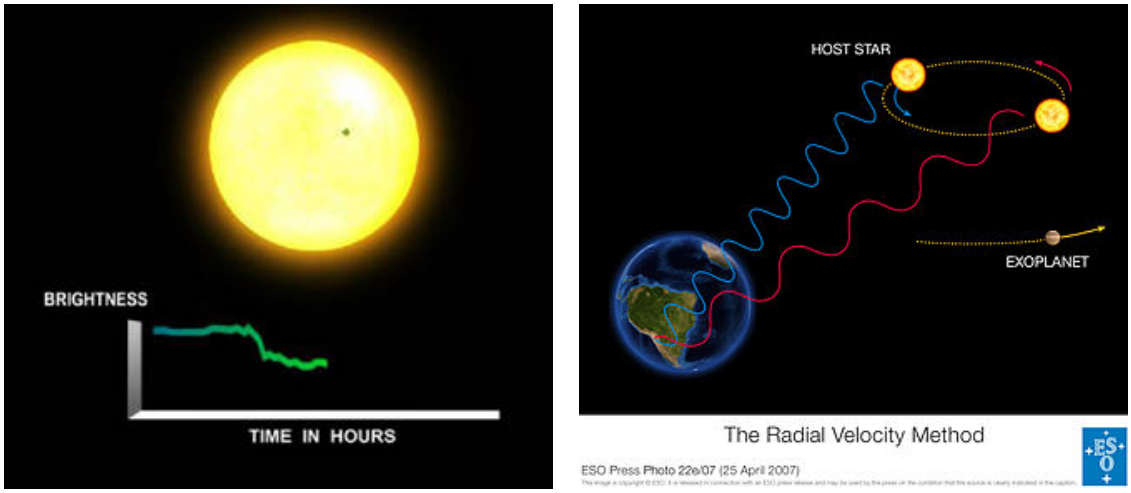
Currently Kepler accounts for the majority of exoplanet discoveries using this method but future missions such as TESS [2], and PLATO 2.0 [3] could surpass this.

2) Radial Velocity: 2 (b) accounted for 18.83% of all planet discoveries. Exoplanets are detected using radial velocity method by observing the effect of the exoplanets gravitational pull on a star. When a star moves towards us, its spectrum is blue shifted and when it moves away from us its spectrum is red shifted. Observing these shifts regularly can determine if the movement is due to an exoplanet and in turn determine that exoplanets mass. Ground based observatories such as La Silla Observatory and W.M.Keck observatory are mainly responsible for these types of discovery.

3) Other methods: 4.23% of planets discovered were found using various techniques with names like as Microlensing and Imaging.

Transit method provide the dimensions of a planet and radial method provide mass of a planet. Using both measurements together it is possible to calculate the density of the planet.

In this report the **literary review** II describes papers which highlight the need for machine learning in astronomy. Kepler data was chosen for testing because investigations revealed that most planets were discovered by this mission using the transit method and the paper "Big Universe, Big Data: Machine Learning and Image Analysis for Astronomy" [4] also described Kepler as a good candidate for testing machine learning algorithms.The machine learning algorithms applied by NASA and google were reviewed. The deep learning technique applied by google on publicly available Kepler data resulted in the discovery of two planets overlooked by NASA. This google code was recently publicly released and may provide a good source for further investigations. The review also looks at the short history of exoplanet discovery and its current state.



(a) (b)

Fig. 2. (a) Transit (When a planet crosses in front of its star) / (b) Radial Velocity - Examine red shift and blue shift

The section **research question** III then poses the question, whether it possible to find other exoplanets overlooked by NASA. The report report continues with...

- Proposed Approach IV
- Proposed Implementation V
- Proposed Evaluation VI
- Project Plan VII

II. Literature Review

The astronomical community has highlighted the need for machine learning due to the ever increasing amount of data which is being gathered. Currently the Chinese MingantU SpEctral Radioheliograph (Muser) records about 100TB of raw data per month which prompted the Chinese academy of sciences to note that it has become ”urgent” to find automatic algorithms for processing big data [5]. The paper ”Big Universe, Big Data: Machine Learning and image analysis for astronomy” [4] reported that volumes of entire surveys a decade ago can now be acquired in a single night and in the next few years we will need machine learning systems that can process terabytes of data in near real-time with high accuracy. This paper also refers to the *Kepler exoplanet data* available at the Mikulski Archive for Space Telescopes [6] which is described as a *”valuable dataset for testing detection algorithms”*. It is interesting to note that all 5 authors of this paper are Danish computer scientists (not astronomers).

A visualisation database of Exoplanets is shown in Fig 3. This dynamic infographic showed that transit method accounted for 76.88% of all exoplanet discoveries and Kepler was the most successful mission using this method. This and the paper ”Big Universe, Big Data” [4] indicated Kepler data is a good data choice for testing detection algorithms.

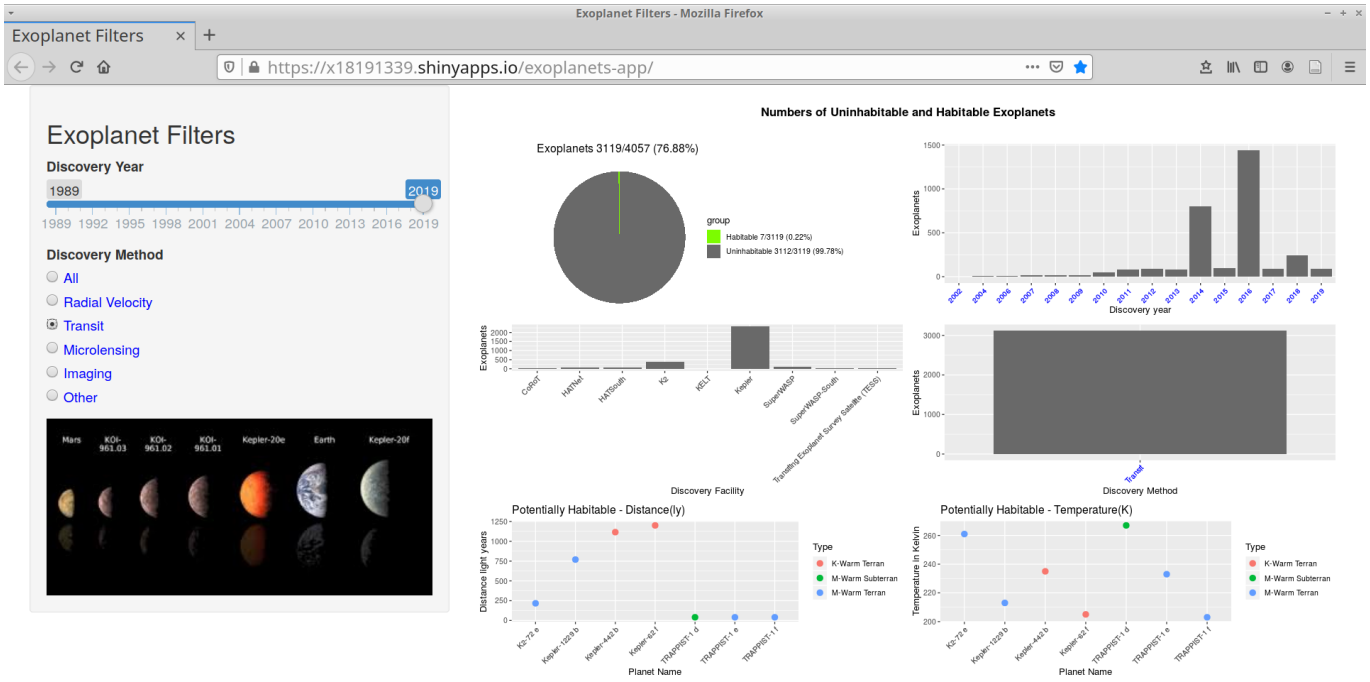


Fig. 3. Exoplanet Discovery Methods https://x18191339.shinyapps.io/exoplanets-app/

The easiest planets to spot are large, nearby and orbit quickly but Kepler was hunting for habitable planets which are generally smaller and harder to spot. The first planet ever discovered is called HD 114762b and it was discovered in 1989 by Latham [7] and confirmed in 1991 in a paper by Cochran [8]. This planet is 11 times more massive than Jupiter, 128.7 light years away and orbits its star in 84 day, but it was initially misclassified as a brown dwarf [1]. The first earth like planet,called GJ66GC, was only discovered in 2011. GJ66GC is a super-Earth, which is larger than Earth, but smaller than Uranus [9]

In 2013 the Los Angeles Times reported there were estimated to be up to 40 billion planets in our galaxy the milky way [10] but this is an underestimate as Kepler results have shown that there are more planets than stars in our galaxies [11]. Our galaxy the milky way has approximately 100 billion stars[12]. The Kepler Space Telescope observed a small portion of the milky way see Fig 1 but it still managed to photometrically observe over 200,000 stars.

McCauliff [13] described how random forest was used for the first time to find Planetary Candidates in the Kepler mission. This was done by transforming transit-like detections into a uniform set of real-numbered attributes called Threshold Crossing Events (TCE's)[2]. The random forest algorithm became a tool called the robovetter which was used to predict whether a TCE

[1] A brown dwarf is a type of sub stellar object occupying the mass range between the heaviest gas giant planets and the lightest stars

[2] A threshold crossing event (TCE) is a sequence of significant, periodic, planet transit-like features in the light curve of a target star

corresponded to a planet or a non-transiting phenomenon such as instrument noise or a passing binary star. They found 3697 planet candidates (PCs) from a set of 18,406 transit-like features detected on more than 200,000 distinct stars. McCauliff also pointed out that such tools would be needed for future planetary missions such as K2 [14],TESS [2], and PLATO 2.0 [3].

After an exoplanet is classified as a potential candidate (PC) it requires independent confirmation using a different method but Kepler's profusion of planet candidates overwhelmed the resources available for ground-based Radial Velocity [15]. Using the robovetter proved useful in helping astronomers wasting time trying to confirm non-existent planets but it required some trade-offs. If not enough vetting is done the amount of false positives could be overwhelming, whereas too much vetting may result in planets being missed and robovetter missed some planets.

Kepler-62f was wrongly classified as a false positive and of the planets discovered to date Kepler-62f is arguably the known exoplanet most likely to be habitable. In the abstract of the paper "Kepler-62f: Kepler's first small planet in the habitable zone, but is it real?" [16] Borucki notes "While exceptionally useful for producing a uniform catalogue, these algorithms sometimes misclassify planet candidates as a false positive".

Two overlooked exoplanets were later found by a google engineer Chris Shallue. In 2018 Shallue [17] used deep learning, specifically Convolutional Neural Network (CNN), to identify several new planet candidates which were at the limit of the Kepler mission's detection sensitivity. Two of these candidates were later confirmed as planets. Shallue only looked at 670 of the 200,000 light curves which suggests there may still be planetary candidates waiting to be discovered.

Other astronomers are also turning to CNN. In October 2019 Chausev et al.[18] reported in the Royal Astronomical Society how they applied CNN to the problem of exoplanet vetting. Using this method they reduced the time required by manual vetting by half and recovered 13 out of 14 confirmed planets observed by the ground based Next-Generation Transit Survey (NGTS).

These recent publications on using CNN for discovering exoplanets lead to the research question which is discussed below.

III. Research Questions

Google released the code which was used to discover two new exoplanets [17] and to quote the engineer who made the discoveries in a google blog[19]
"We've only searched 670 stars out of 200,000 observed by Kepler — who knows what we might find when we turn our technique to the entire dataset"
. He went on to explain that the code did not work as well as Robovetter system at rejecting certain types of simulated false positives. This begs the (research) question can we improve algorithms used to predict exoplanets. There are several ways to address this question.

- Can the accuracy of convolutional neural network (CNN) method used by Shallue 4 be improved.
- Can we improve planet prediction using other methods like Tree-CNN [20]
- Can the speed of prediction be improved using big data solutions like spark.

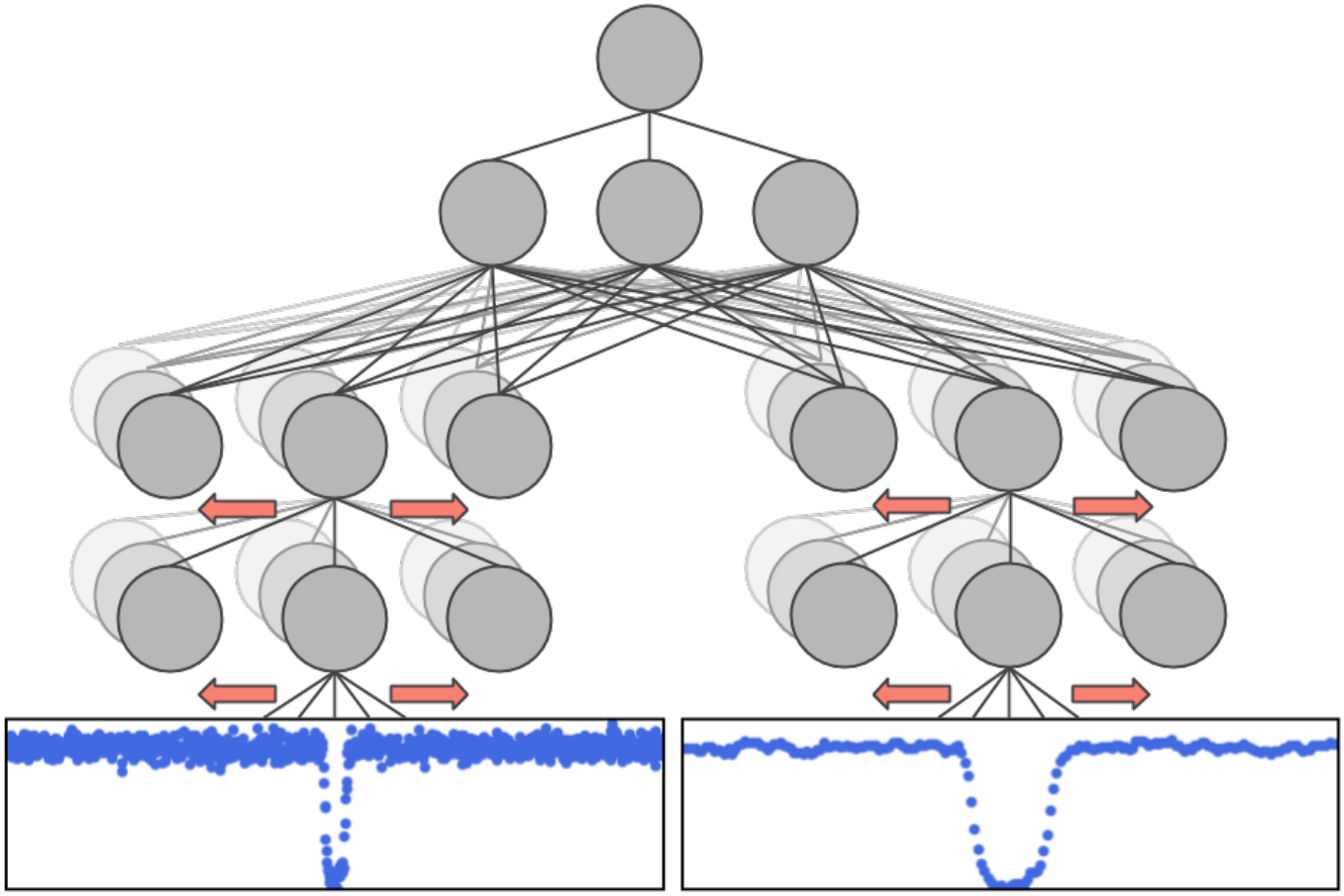


Fig. 4. Deep Learning

IV. Proposed Approach

The CNN software used by Shallue to predict exoplanets was released as an open source package called exoplanet-ml 5. This software will be used as the research starting point. The following approach will be followed...

- Set up the software environment. Maybe spark on top of hadoop.
- Download and install exoplanet-ml software Fig 5 and test
- Modify software to run on spark and test if it improves speed.
- Create an ROC comparing the prediction accuracy reported using random forest [13] against that using CNN [17]
- Try different algorithms e.g. Tree-CNN to see if they offer any improvement
- As Shallue has only looked at 670 of 200,000 TCE's try to identify the best curves to search for new planets.
- Can the software be applied to other problems in astronomy?

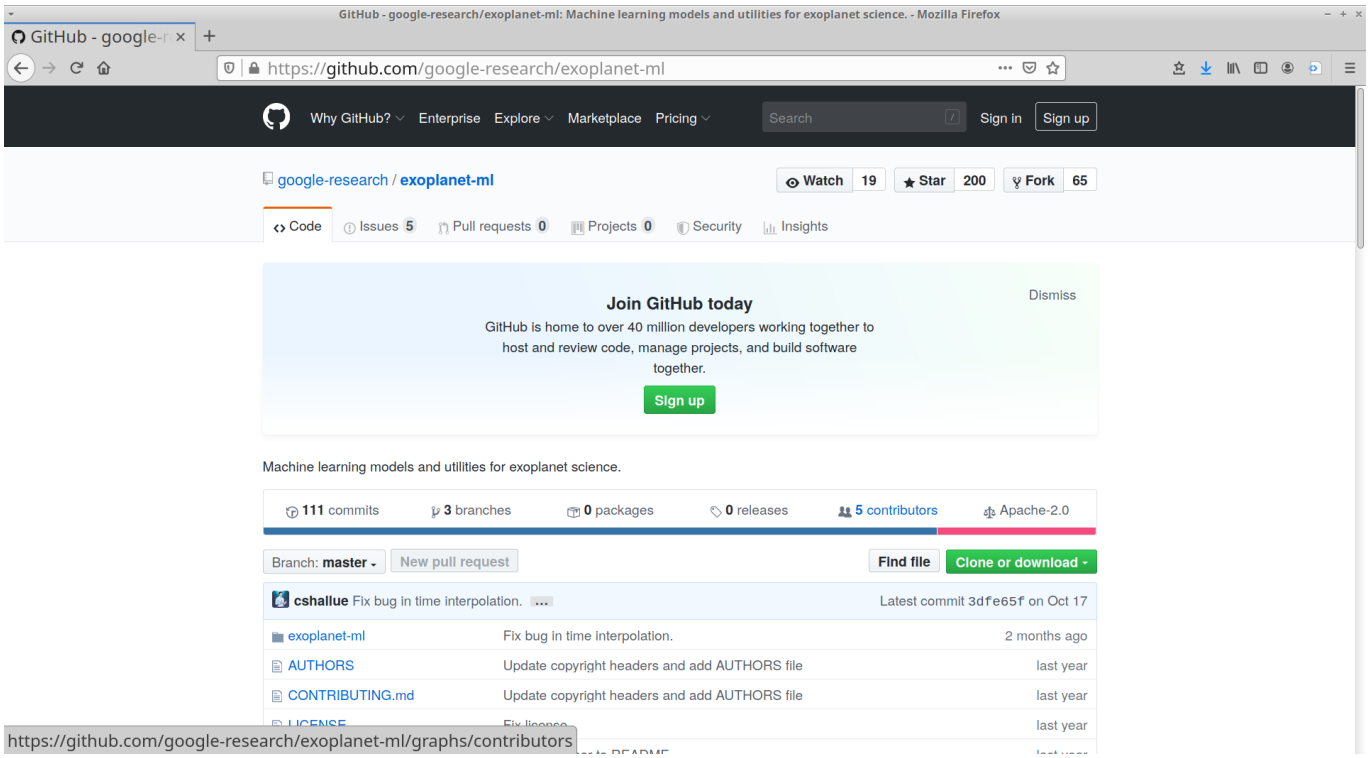


Fig. 5. Google open source exoplanet soft https://github.com/google-research/exoplanet-ml

V. Proposed Implementation

The data to be used in the implementation is publicly available at two sites …

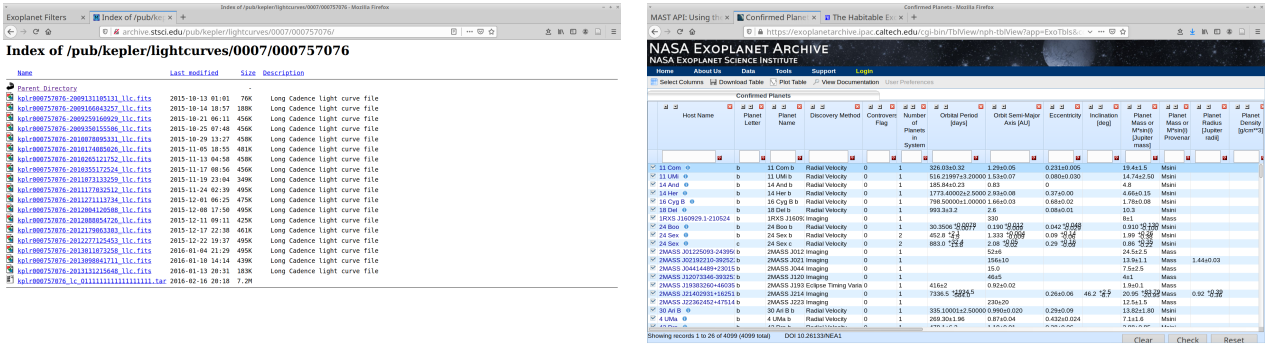
- The Kepler data light curves are publicly available through the Mikulski Archive for Space Telescopes (MAST) 6 (a)
- The TCE's for confirmed planets are at the NASA exoplanet catalogue 6 (b).

The data from MAST can be up to **90GB** so a big data solution is needed. A software pipeline will be needed and several scripts will be created to download and install the data locally. Following this a script will be needed to import the data into a database and maybe convert it from fits [3] format to csv format.

The exoplanet-ml software will then be downloaded and tested from the site shown in Fig 5. Modification of the software may be needed but confirmation of correct installation can be achieved when all the unit tests associated with this software pass.

When the software works it can then be modified as required. To aid software development tools like jupyter labs and git will be installed and configured.

Software modifications will be made to follow the proposals provided in section IV. It will also be possible to avail of the most recent software packages and machine learning algorithms, although this may require tweaking the software.



(a) (b)

Fig. 6. (a) MAST catalog http://archive.stsci.edu/pub/kepler/lightcurves/0062/006273239/ (b) NASA catalogue https://exoplanets.nasa.gov/

[3] fits is the open astronomy format used to store the light curves

VI. Proposed Evaluation

The time taken to run each algorithm can be checked by recording the time at the start and end of each script.
The receiver operating characteristic curve (ROC) will be used to compare the accuracy of each algorithm. Using this method we can find the Area Under Curve which gives us a number that can be used to compare the performance of the different machine learning algorithms.
McCauliff [13] used and ROC when comparing the performance of random forest, Naïve Bayes and K-NN as the best methods for predicting exoplanets 7. The error rate for classifying exoplanets candidates using random forest was reported as 2.81% which was the lowest and hence the method this was chosen for the NASA robovetter.
The same method will be used to check the results of CNN utilized by Shallue [17] against the random tree results. Other machine learning algorithms can also be compared using this method.

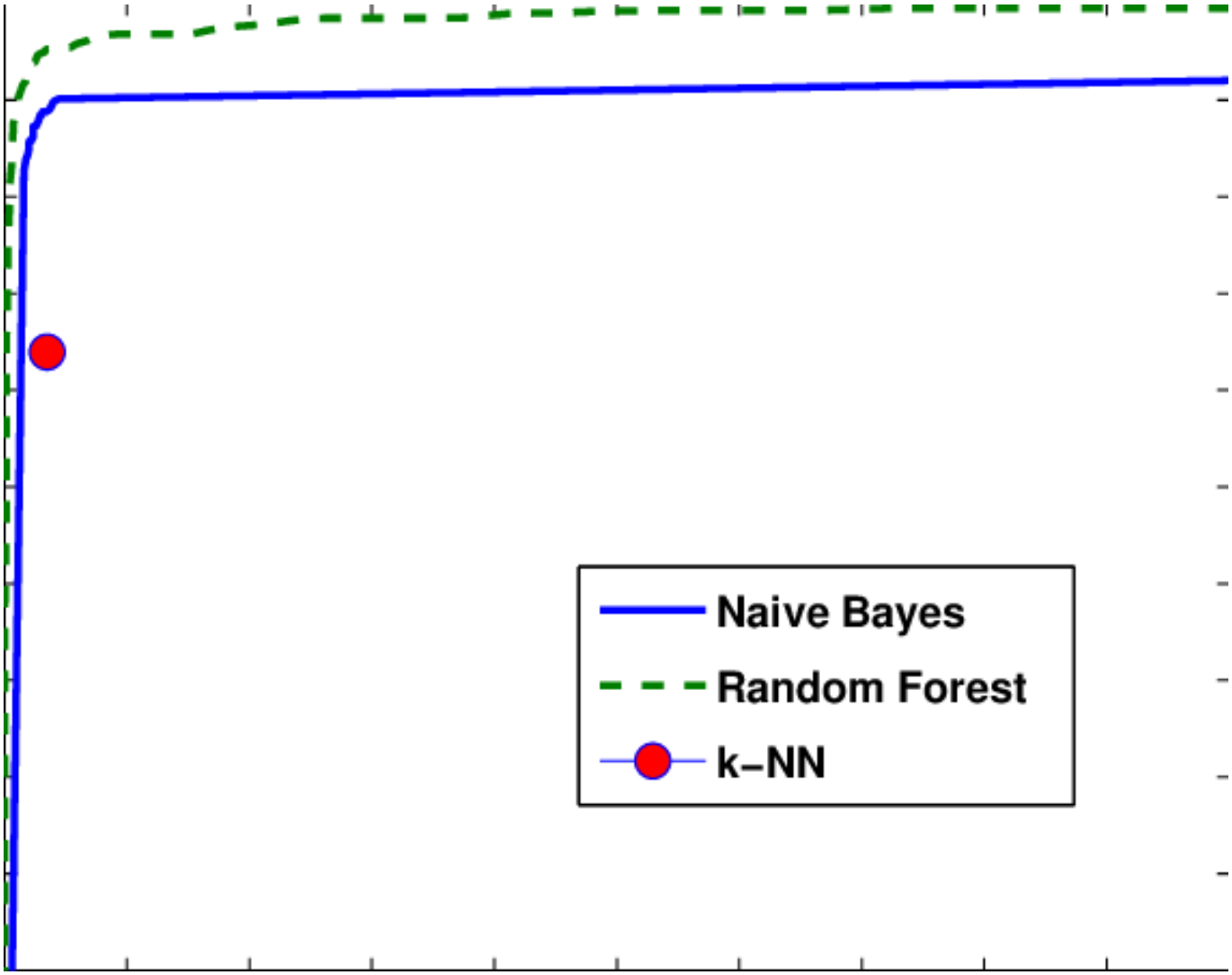


Fig. 7. ROC curve for Kepler's robovetter. (K-NN is a spot when K=1)

VII. Project Plan

A Gantt chart of the expected tasks is show in figure VII. The project is foreseen to last approximately 13 weeks from 01/02/2019 until 30/04/2019.

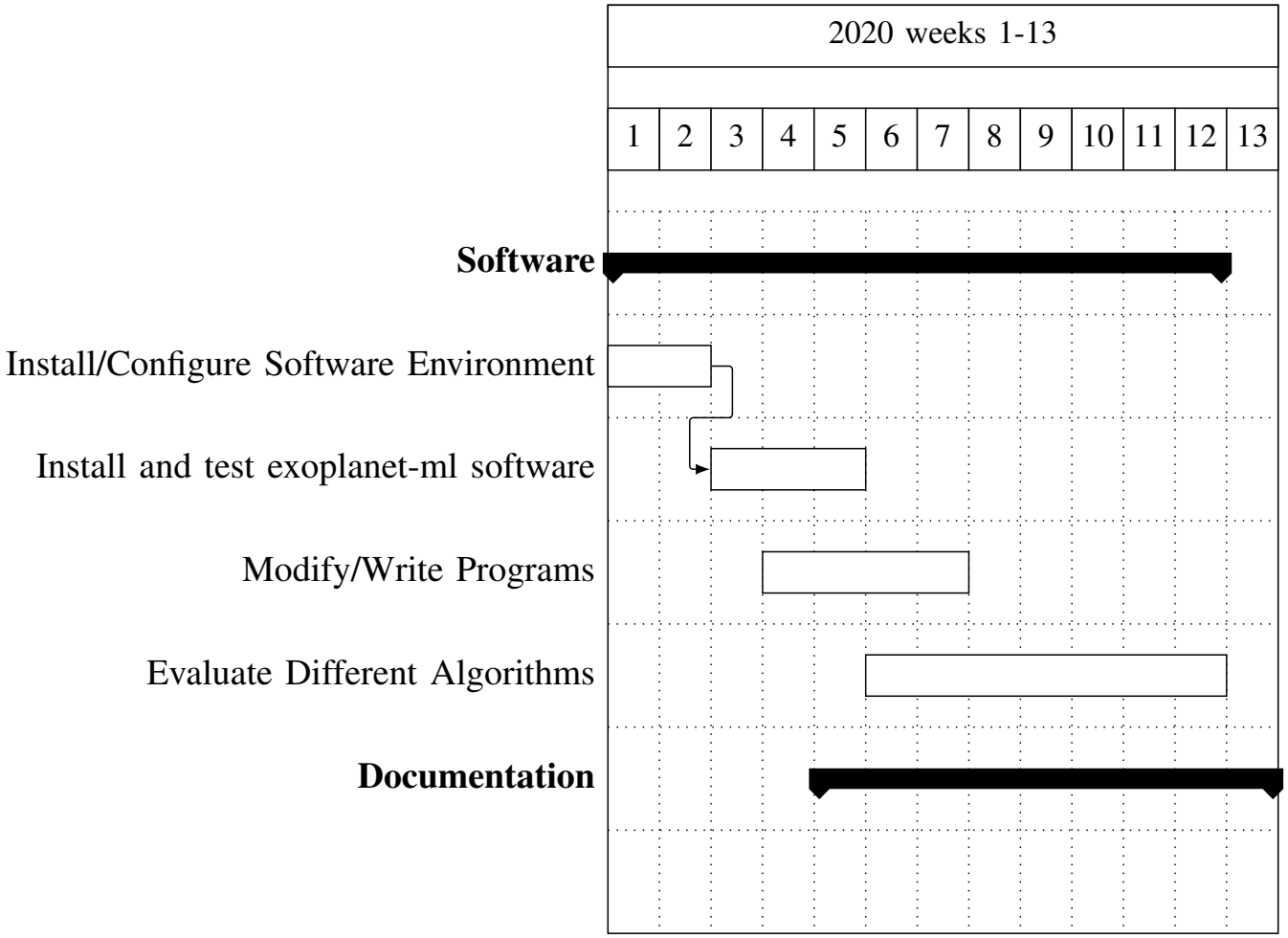


Fig. 8. Gantt chart: 01/02/2019 until 30/04/2019

References

[1] Planetary Society. *Radial Velocity*. URL: https://www.planetary.org/explore/space-topics/exoplanets/radial-velocity.html (visited on 11/16/2019).

[2] George R. Ricker et al. "The Transiting Exoplanet Survey Satellite". In: *Journal of Astronomical Telescopes, Instruments, and Systems* 1.1 (Oct. 24, 2014), p. 014003. ISSN: 2329-4124. DOI: 10.1117/1.JATIS.1.1.014003. arXiv: 1406.0151. URL: http://arxiv.org/abs/1406.0151 (visited on 11/27/2019).

[3] H. Rauer et al. "The PLATO 2.0 mission". In: *Experimental Astronomy* 38.1 (Nov. 1, 2014), pp. 249–330. ISSN: 1572-9508. DOI: 10.1007/s10686-014-9383-4. URL: https://doi.org/10.1007/s10686-014-9383-4 (visited on 11/27/2019).

[4] Kremer, J. et al. "Big Universe, Big Data: Machine Learning and Image Analysis for Astronomy". In: *IEEE Intelligent Systems, Intelligent Systems, IEEE, IEEE Intell. Syst.* 2 (2017), p. 16. ISSN: 1541-1672. DOI: 10.1109/MIS.2017.40.

[5] Monica Young. "Machines Learning Astronomy: The new era of artificial intelligence & Big Data is changing how we do astronomy". In: *Sky & Telescope* 134.6 (Dec. 2017), pp. 20–27. ISSN: 00376604. URL: http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,cookie,shib&db=a9h&AN=125472912&site=eds-live&scope=site&custid=ncirlib (visited on 10/15/2019).

[6] MAST. *Kepler*. MAST. URL: https://archive.stsci.edu/kepler/ (visited on 11/24/2019).

[7] David W. Latham et al. "The unseen companion of HD114762: a probable brown dwarf". In: *Nature* 339.6219 (May 1989), pp. 38–40. ISSN: 1476-4687. DOI: 10.1038/339038a0. URL: https://www.nature.com/articles/339038a0 (visited on 12/01/2019).

[8] William D. Cochran, Artie P. Hatzes, and Terry J. Hancock. "Constraints on the companion object to HD 114762". In: *Astrophysical Journal* 380.1 (Oct. 10, 1991). ISSN: 0004-637X. DOI: 10.1086/186167. URL: https://researchers.dellmed.utexas.edu/en/publications/constraints-on-the-companion-object-to-hd-114762 (visited on 11/25/2019).

[9] NASA. *GJ 667 C c*. 2019. URL: https://exoplanetarchive.ipac.caltech.edu/cgi-bin/DisplayOverview/nph-DisplayOverview?objname=GJ+667+C+c (visited on 12/14/2019).

[10] Khan. *Milky Way may host billions of Earth-size planets*. Los Angeles Times. Nov. 5, 2013. URL: https://www.latimes.com/science/la-sci-earth-like-planets-20131105-story.html (visited on 10/14/2019).

[11] Dennis Overbye. "Kepler, the Little NASA Spacecraft That Could, No Longer Can". In: *The New York Times* (Oct. 30, 2018). ISSN: 0362-4331. URL: https://www.nytimes.com/2018/10/30/science/nasa-kepler-exoplanet.html (visited on 12/13/2019).

[12] ESA. *How many stars are there in the Universe?* URL: https://www.esa.int/Science_Exploration/Space_Science/Herschel/How_many_stars_are_there_in_the_Universe (visited on 11/16/2019).

[13] Sean D. McCauliff et al. "Automatic Classification of Kepler Planetary Transit Candidates". In: *The Astrophysical Journal* 806.1 (June 3, 2015), p. 6. ISSN: 1538-4357. DOI: 10.1088/0004-637X/806/1/6. arXiv: 1408.1496. URL: http://arxiv.org/abs/1408.1496 (visited on 12/14/2019).

[14] Steve B. Howell et al. "The K2 Mission: Characterization and Early results". In: *Publications of the Astronomical Society of the Pacific* 126.938 (Apr. 2014), pp. 398–408. ISSN: 00046280, 15383873. DOI: 10.1086/676406. arXiv: 1402.5163. URL: http://arxiv.org/abs/1402.5163 (visited on 11/27/2019).

[15] Jack J. Lissauer and Joann Eisberg. "New Astronomy Reviews special issue: History of Kepler's major exoplanet "firsts"". In: *New Astronomy Reviews* 83 (Nov. 1, 2018), pp. 1–4. ISSN: 1387-6473. DOI: 10.1016/j.newar.2019.04.002. URL: http://www.sciencedirect.com/science/article/pii/S1387647319300144 (visited on 11/27/2019).

[16] William Borucki et al. "Kepler-62f: Kepler's first small planet in the habitable zone, but is it real?" In: *New Astronomy Reviews* 83 (Nov. 1, 2018), pp. 28–36. ISSN: 1387-6473. DOI: 10.1016/j.newar.2019.03.002.

[17] Christopher J. Shallue and Andrew Vanderburg. "Identifying Exoplanets with Deep Learning: A Five-planet Resonant Chain around Kepler-80 and an Eighth Planet around Kepler-90". In: *Astronomical Journal* 155.2 (Feb. 2018), p. 1. ISSN: 00046256.

[18] Alexander Chaushev et al. "Classifying exoplanet candidates with convolutional neural networks: application to the Next Generation Transit Survey". In: *Monthly Notices of the Royal Astronomical Society* 488.4 (Oct. 2019), p. 5232. ISSN: 00358711.

[19] Chris Shallue. *Open Sourcing the Hunt for Exoplanets*. Google AI Blog. Aug. 3, 2018. URL: http://ai.googleblog.com/2018/03/open-sourcing-hunt-for-exoplanets.html (visited on 11/29/2019).

[20] Deboleena Roy, Priyadarshini Panda, and Kaushik Roy. "Tree-CNN: A hierarchical Deep Convolutional Neural Network for incremental learning". In: *Neural Networks* 121 (Jan. 1, 2020), pp. 148–160. ISSN: 0893-6080. DOI: 10.1016/j.neunet.2019.09.010. URL: http://www.sciencedirect.com/science/article/pii/S0893608019302710 (visited on 12/14/2019).