Identifying Potentially Habitable Exoplanets Using Machine Learning: A Broad Overview

Benjamin Bartek
Bellevue University
Bellevue, NE 68005, USA
bbartek@my365.bellevue.edu

Abstract

Since the dawn of recorded history, humanity has looked to the stars with a persistent question: are we alone in the universe? [22] Now well into the space age, 21st century humanity has seen breakthroughs in space and machine learning technology that are crucial to the search for life beyond our Solar System. Data sources such as the Habitable Exoplanets Catalog and data collected from the Kepler space observatory have been a boon to the search for potentially habitable exoplanets. Kepler data helps identify foreign planetary bodies by identifying dips in light output as exoplanets transit the path of stars located beyond our Solar System. [13, 15] Scientists can then use more targeted techniques to collect additional habitability data such as proximity to the nearest star, surface water liquidity, planetary composition, and more. [5, 11] This abstract summarizes the role of machine learning in the discovery of potentially habitable exoplanets.

Author Keywords

Stars; Exoplanets; Planets; NASA; Data; Kepler; Machine Learning; Neural Networks; Habitable; Habitability; Transit; Threshold-Crossing, Event.

ACM Classification Keywords

I.2 ARTIFICIAL INTELLIGENCE

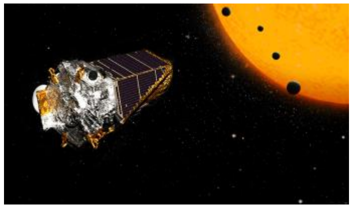


Figure 1: An artistic rendering of the Kepler space observatory. Image CC by NASA/JPL-CalTech at jpl.nasa.gov

Figure 2: An artistic depiction of planetary system K2-138. Image CC by NASA/JPL-CalTech at jpl.nasa.gov

Introduction

The search for extraterrestrial life is a nascent scientific discipline that has emerged concomitantly with technological advancements in computing and optics that were designed for space observation and exploration. Advancements in statistical data analysis - and machine learning in particular - have also enabled scientists to analyze large volumes of data more efficiently and accurately than ever before. Using machine learning algorithms to analyze light data collected from nearby stars has resulted in the identification of likely and confirmed exoplanets of both the uninhabitable and potentially habitable varieties. This abstract synthesizes the author's review of recent scientific literature to provide a non-exhaustive summary of the common data sources, habitability factors, and machine learning techniques used to identify potentially habitable exoplanets.

Context and Limitations

The existence of life on Earth is axiomatic to all but a certain subset of philosophers. But while it is increasingly probable that life exists elsewhere in a universe comprised of trillions of stars, Earth is so far the only planet confirmed to harbor life in either a simple or complex form. [11, 21] The search for extraterrestrial life even within our own Solar System continues, driven by humanity's desire to observe and understand previously undiscovered properties of the universe. The following discussion, however, relates solely to the identification of exoplanets located beyond Earth's Solar System.

Data Sets

To analyze data, one must first obtain data sets. While many scientific agencies are tasked in part with the evidence-based search for extraterrestrial life, humans have only recently developed the tools necessary to search for life beyond our Solar System. [4, 8] One such tool is NASA's Kepler space observatory. Launched in 2009, the Kepler and K2 Missions have observed more than 500,000 stars in the nearby galactic neighborhood, using a photometer to detect variations in light as planets transit the observed stars. [8, 15, 17, 19] Many of the observed stars have temperature and surface gravity measurements similar in range to those of our own Sun. [7]

Large volumes of data from the Kepler Missions are publicly available in three primary sources: 1) the NASA Exoplanet Archive, 2) NASA Kepler Mission Data housed at the Barbara A. Mikulski Archive for Space Telescopes, and 3) the University of Puerto Rico at Arecibo Habitable Exoplanets Catalog. [17, 18, 25] The NASA Exoplanet Archive "…includes interactive tables containing properties of all published exoplanets, Kepler planet candidates, threshold-crossing events, data validation reports and target stellar parameters, light curves from the Kepler and CoRoT missions and from several ground-based surveys, and spectra and radial velocity measurements from the literature". [1]

Planetary Candidate Identification

Planetary candidates are identified using the transit method, searching for dips in light as a planet transits across the face of a star. [1, 13] Light curves are analyzed over time to determine how often the transit occurs. [1] "Each transit-like event detected by the pipeline with a signal-to-noise ratio greater than 7 constitutes a threshold-crossing event (TCE). These TCEs are further studied and characterized to identify planet candidates, eclipsing binaries, and false

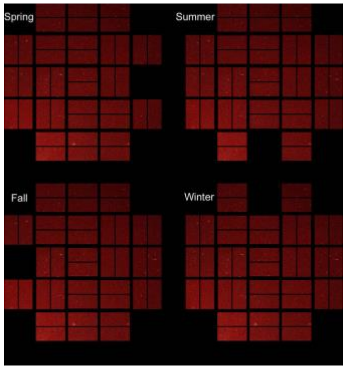


Figure 3: Kepler fields of view vary by season. Image CC by nasa.gov

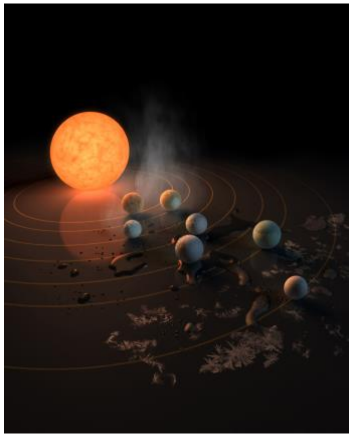


Figure 4: An artistic rendering of the ultra-cool dwarf star TRAPPIST-1, with seven Earth-size planets in orbit. Image CC by NASA/JPL-CalTech at jpl.nasa.gov

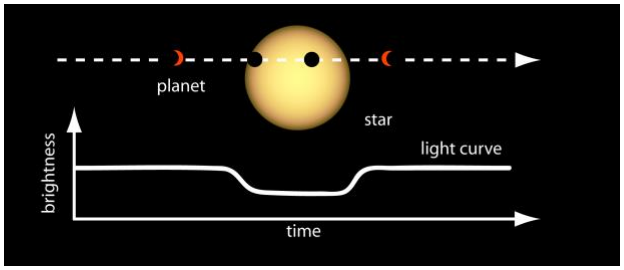


Figure 5: Kepler's photometer dips in light as an exoplanet transits the face of its host star. Image CC by nasa.gov

positives. The remaining objects are placed on the Kepler Object of Interest (KOI) list and are subjected to follow-up observations and further analysis to confirm or validate their planetary status." [1] Planetary validation analysis also includes other criteria such as a "…(1) … mass (or minimum mass) estimate that is equal to or less than 30 Jupiter masses, (2) the properties of the planet are described in the peer-reviewed literature, and (3) sufficient follow-up observations and validation have been undertaken to deem the possibility of the object being a false positive as unlikely."[1]

Habitability Factors

Once a planetary candidate is confirmed, a wide array of additional factors determine the potential habitability of a given exoplanet. The simplest measurement (though far from a guarantee of actual habitability) is whether the exoplanet orbits within the habitable zone (HZ) of its host star at a distance relative to the heat and mass of the star that allows the exoplanet to retain surface water. [12, 14, 20, 22]

More complex measurements exist as well. The Earth Similarity Index (ESI) is one measure of habitability, assigning an ESI score to exoplanets based on the similarity of the exoplanet's trajectory, density, temperature, and radius compared to Earth's measurements. [2] Similar indices include the Biological Complexity Index (BCI) and the Planetary Habitability Index (PHI). [3] The PHI considers planetary composition factors such as "…a stable and protected substrate, energy, appropriate chemistry and a liquid medium[,]" and the BCI adds temperature, geological complexity, and age as factors. [5, 11] Yet another significant factor for the development of life is the existence of a magnetosphere, which holds an atmosphere in place and shields the planet from harmful radiation. [2]

Astute observers would note that most habitability factors are not observed by the Kepler Missions. Indeed, more targeted ground-based surveys are necessary to capture additional observational data pertaining to habitability. [1, 9, 15] Moreover, because humanity has only observed Earth-based life forms, there is still much to discover regarding the exact parameters required for an exoplanet to host life. Alternative life forms could conceivably thrive in a methane environment rather than a water environment, for example. [21]

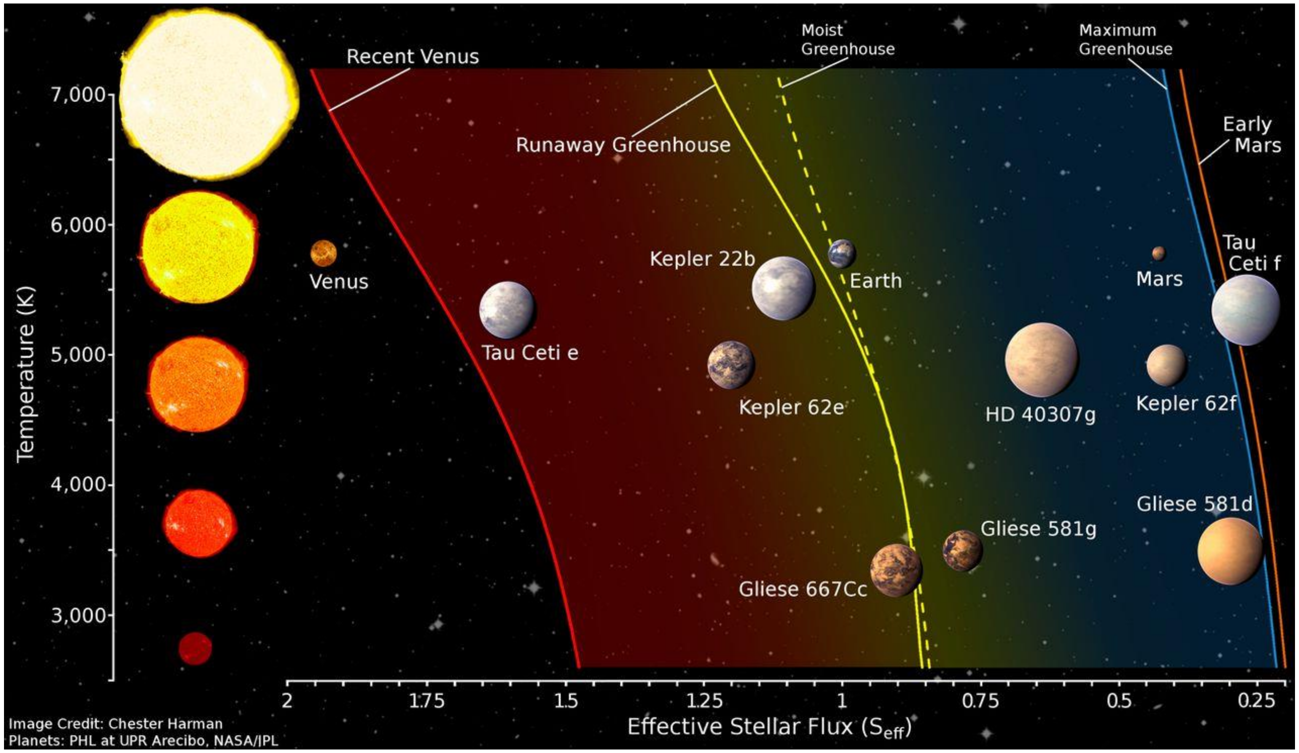


Figure 6: "Diagram showing different HZ boundaries for stars ranging in spectral type from F0 to M7. Various planets within our Solar System are shown, along with selected exoplanets." [12]

Machine Learning Methods

The cited literature mentions little in the way of data cleaning and preparation, but one type of calibration is key early in the data science process. Raw photometric light curve data must have pixels calibrated to remove the negative "…effects of cosmic rays and…bias, dark current, gain, etc." [19] The light curve data must also undergo radial and temporal data validation to be useful for identifying Kepler Objects of Interest. [19] One light curve study determined that half-sibling

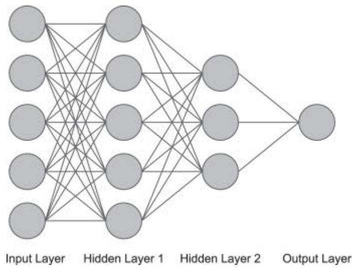


Figure 7: A fully connected neural network. Data is inserted into the input layer, with the output layer representing the predictive output of hidden layers of interconnected algorithms with artificial learning features. [24]

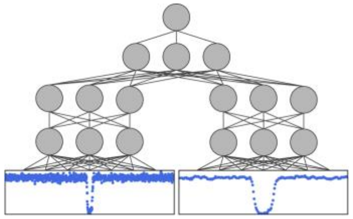


Figure 8: "Fully connected neural network architecture for classifying light curves, with both global and local input views." [24]

regression is an effective method of removing such confounders. [23]

In the machine learning context, habitability factors can be quantified as weighted variables in habitability index scores. One basic algorithm for calculating the Biological Complexity Index is as follows:

$$BCI_{abs} = (S \times E \times T \times G \times A)^{1/5}$$

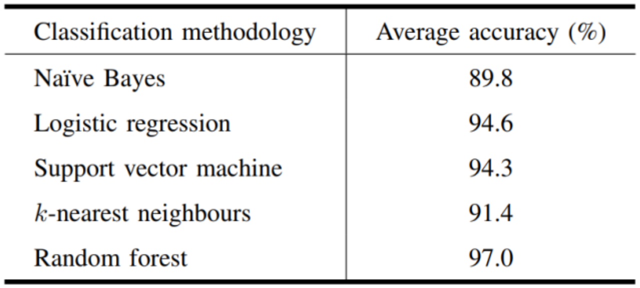
Here, S represents substrate, E represents energy, G represents geological complexity, and A represents age. [11]

Among the variety of methods available for analyzing data for planetary identification, the following algorithms and classifiers are predominant:

- Deep Convolutional Neural Networks [24]
- Recursive Feature Elimination [24]
- Support Vector Machines [19, 24]
- Gaussian Naïve Bayes [3, 19]
- K-Nearest Neighbors [3, 5, 19]
- Random Forest/Decision Trees [3, 16, 19]

These algorithms and classifiers are commonly and reliably used in many other disciplines as well. Algorithm selection depends largely on the intended goal, the selected variables, and the accuracy or fit of the model. Gaussian Naïve Bayes is a probabilistic classifier, K-Nearest Neighbors is a simple instance-based classifier, support vector machines are hard-boundary classifiers, and deep convolutional neural networks layer algorithms comprised of trained, weighted variables. [3, 5, 19, 24] Random forests are "…an ensemble of decision trees. Each tree votes for the classification of an unknown object based on a vector of attributes…that describe the object in question. The classification the forest assigns to an object is the plurality vote of the random forest." [16] Trained random forest decision outputs are particularly useful for KOI identification, with an error rate of only 3 percent in one peer reviewed study. [19]

In another frequently cited study, a trained deep convolutional neural network predicted with 98.8 percent accuracy whether a light curve was indicative of a planetary body or simply a false positive due to confounders. [24] As a result, the study identified two new planets with statistical certainty. [24] That study is but one of many that have identified new exoplanets.



Classification methodology	Average accuracy (%)
Naïve Bayes	89.8
Logistic regression	94.6
Support vector machine	94.3
k-nearest neighbours	91.4
Random forest	97.0

Figure 9: Machine learning accuracy rates for KOI according to studies by Nigri and Arandjelovic. [19]

Results

The application of machine learning methods to Kepler data has yielded remarkable results:

- 3,791 Confirmed Exoplanets [18]
- 14 Potentially Habitable Exoplanets (Conservative) [25]

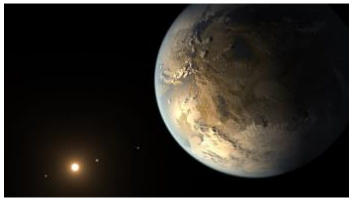


Figure 10: An artistic rendering of Kepler 186-f, the first Earth-size planet confirmed to orbit within the habitable zone of another star. Image CC by nasa.gov

- 41 Potentially Habitable Exoplanets (Optimistic) [25]

These conclusions should be tempered by the receipt of brand new data from the European Space Agency's Gaia spacecraft that indicates that fewer planets may actually orbit within the HZ than presently identified. [26] Nevertheless, future data and analysis should yield even more dividends, and past results can be used as training data for new research. [20, 24, 25] Though the Kepler spacecraft retired in October 2018, the James Webb Space Telescope, with more precise instruments for observing the exoplanet atmospheres directly, is scheduled to launch in 2021. [6, 10]

Conclusion

NASA's Kepler Missions have provided an unprecedented amount of raw photometry data for use by both scientists and amateurs. Machine learning is an invaluable array of tools that, with the design and insight of humans, analyzes and classifies data to identify exoplanets. Properly fitted algorithms and models yield real advances in the identification off potentially habitable worlds beyond our Solar System. The search for new worlds is in its infancy, with newly discovered planets and insights on the horizon.

Acknowledgements

Special thanks to Dr. Nicolas LaLone, Bellevue University, for his expertise and guidance in the field of Data Science. My gratitude also goes out to all of my classmates, with whom I share this journey to become a data scientist. Finally, none of this research would have been possible without NASA, the researchers who have spent countless hours poring over data worldwide, and the inspiration of dreamers and astronomers such as Galileo, Copernicus, Neil deGrasse Tyson, and Sir Isaac Newton.

References

1. R.L. Akeson, X. Chen, D. Ciardi, et al. 2013. The NASA exoplanet archive: data and tools for exoplanet research. *Publications of the Astronomical Society of the Pacific* 125, 930 (July 2013), 23 pages. ArXiv: 1307.2944v1. Retrieved from https://arxiv.org/pdf/1307.2944.pdf
2. D.J. Armstrong, C.E. Pugh, A.-M.Broomhall, D.J.A. Brown, M.N.Lund, H.P. Osborn, and D.L. Pollacco. 2016. The host stars of Kepler's habitable exoplanets: superflares, rotation and activity. *Monthly Notices of the Royal Astronomical Society* 455, 3 (Jan. 2016), 3110-3125. DOI: https://doi.org/10.1093/mnras/stv2419
3. Suryoday Basak, Surbhi Agrawal, Snehanshu Saha, Jeremiel Theophilus, Kakoli Bora, Gouri Deshpande, and Jayant Murthy. 2018. Habitability classification of exoplanets: a machine learning insight. (May 2018). arXiv:1805.08810v1. Retrieved from https://arxiv.org/pdf/1805.08810.pdf
4. Natalie M. Batalha. 2014. Exploring exoplanet populations with NASAs Kepler Mission. In *Proceedings of the National Academy of Sciences* 111, 35 (July 2014), 12647–12654. DOI:http://dx.doi.org/10.1073/pnas.1304196111
5. Kakoli Bora, Snehanshu Saha, Surbhi Agrawal, Margarita Safonova, Swati Routh, and Anand Narasimhamurthy. 2016. CD-HP: New habitability score via data analytic modeling. *Astronomy and Computing* 17 (April 2016), 129–143. DOI:https://doi.org/10.1016/j.ascom.2016.08.001
6. Justin Davenport. 2018. Kepler – NASA's planet-hunting spacecraft – retired after running out of fuel. (Oct. 2018). Retrieved October 31, 2018 from https://www.nasaspaceflight.com/2018/10/kepler-retired-after-running-out-of-fuel/

7. Daniel Foreman-Mackey, David W. Hogg, and Timothy D. Morton. 2014. Exoplanet population inference and the abundance of Earth analogs from noisy, incomplete catalogs. *The Astrophysical Journal* 795, 64 (Oct. 2014), 12 pages. DOI:https://doi.org/DOI:10.1088/0004-637X/795/1/64

8. Olivier Guyon. 2017. Habitable exoplanets detection: overview of challenges and current state-of-the-art [Invited]. *Optics Express* 25, 23 (Nov. 2017), 13 pages. DOI:https://doi.org/10.1364/OE.25.028825

9. Kevin Heng. 2017. A new window on alien atmospheres. *American Scientist* 105, 2 (March/April 2017), 86–89. DOI:http://dx.doi.org/10.1511/2017.105.2.86

10. Natalie Hinkel. 2018. Big data on exoplanet composition. *American Scientist* 106, 5 (September-October 2018), 309. DOI:http://dx.doi.org/10.1511/2018.106.5.309

11. Louis N. Irwin, Abel Méndez, Alberto G. Fairén, and Dirk Schulze-Makuch. 2014. Assessing the possibility of biological complexity on other worlds, with an estimate of the occurrence of complex life in the milky way galaxy. *Challenges* 5, 1 (May 2014), 159–174. DOI:https://doi.org/10.3390/challe5010159

12. James F. Kasting, Ravikumar Kopparapu, Ramses M. Ramirez, and Chester E. Harman. 2013. Remote life-detection criteria, habitable zone boundaries, and the frequency of Earth-like planets around M and late K stars. In *Proceedings of the National Academy of Sciences* 111, 35 (October 2013), 12641–12646. DOI:http://dx.doi.org/10.1073/pnas.1309107110

13. Jack J. Lissauer, Rebekah I. Dawson, and Scott Tremaine. 2014. Advances in exoplanet science from Kepler. *Nature* 513, 7518 (Sept. 2014), 336–344. DOI:http://dx.doi.org/10.1038/nature13781

14. Michael C. Lopresto and Hector Ochoa. 2017. Searching for potentially habitable extra solar planets: a directed-study using real data from the NASA Kepler-Mission. *Physics Education* 52, 6 (Sept. 2017), 9 pages. DOI:http://dx.doi.org/10.1088/1361-6552/aa8740

15. Nikku Madhusudhan, Marcelino Agúndez, Julianne I. Moses, and Yongyun Hu. 2016. Exoplanetary atmospheres—chemistry, formation conditions, and habitability. *Space Sciences Series of ISSI From Disks to Planets* 205, 1-4 (Dec. 2016), 285–348. DOI:http://dx.doi.org/10.1007/s11214-016-0254-3

16. Sean D. McCauliff, Jon M. Jenkins, Joseph Catanzarite, et al. 2015. Automatic classification of Kepler planetary transit candidates. *The Astrophysical Journal* 806, 1 (June 2015), 13 pages. DOI:10.1088/0004-637X/806/1/6

17. NASA. 2018. Kepler Data Search and Retrieval. *Barbara A. Mikulski Archive for Space Telescopes*. Retrieved from http://archive.stsci.edu/kepler/

18. NASA. 2018. NASA Exoplanet Archive. Retrieved from http://exoplanetarchive.ipac.caltech.edu

19. Eduardo Nigri and Ognjen Arandjelovic. 2017. Machine learning based detection of Kepler objects of interest. In *5th International Workshop on Emerging Multimedia Systems and Applications* (Oct. 2017), 6 pages. Retrieved from https://research-repository.st-andrews.ac.uk/bitstream/handle/10023/11659/2017_ICMEW_paper1.pdf?sequence=1

20. Erik A. Petigura, Andrew W. Howard, and Geoffrey W. Marcy. 2013. Prevalence of Earth-size planets orbiting Sun-like stars. In *Proceedings of the National Academy of Sciences of the United States of America 2013*, 110, 48 (November 2013), 19273-19278. DOI: https://doi.org/10.1073/pnas.1319909110

21. Ramses Ramirez. 2018. A more comprehensive habitable zone for finding life on other planets.

Geosciences 8, 280 (July 2018), 48 pages. DOI:http://dx.doi.org/10.3390/geosciences8080280

22. Sara Seager. 2013. Exoplanet Habitability. Science 340, 6132 (May 2013), 577–581. DOI:http://dx.doi.org/10.1126/science.1232226

23. Bernhard Schölkopf, David W. Hogg, Dun Wang, Daniel Foreman-Mackey, Dominik Janzing, Carl-Johann Simon-Gabriel, and Jonas Peters. 2015. Removing systematic errors for exoplanet search via latent causes. In *Proceedings of the 32nd International Conference on Machine Learning* 37 (May 2015) 9 pages. Retrieved from http://proceedings.mlr.press/v37/scholkopf15.pdf

24. Christopher J. Shallue and Andrew Vanderburg. 2018. Identifying exoplanets with deep learning: a five-planet resonant chain around Kepler-80 and an eighth planet around kepler-90*. The Astronomical Journal* 155, 94 (Jan. 2018), 21 pages. Retrieved from http://stacks.iop.org/1538-3881/155/i=2/a=94

25. University of Puerto Rico at Arecibo Habitable Exoplanets Catalog. 2018. Retrieved from http://phl.upr.edu/projects/habitable-exoplanets-catalog

26. Mike Wall. 2018. Number of Habitable Exoplanets Found by NASA’s Kepler May Not Be So High After All. (Oct. 2018). Retrieved October 30, 2018 from https://www.space.com/42275-habitable-exoplanets-kepler-discoveries-revised-by-gaia.html