Project Proposal: A Machine Learning Algorithm to Detect Exoplanet Transits

Anna Zuckerman, Leah Zuckerman, Ashutosh Gandhi, and Andrew Floyd

June 2023

1 Introduction and Motivation

Before the groundbreaking 1992 detection of the planetary system PSR1257 + 12 (Wolszczan and Frail, 1992), the existence of exoplanets (planets that orbit stars other than our sun) was only hypothesized. In the few decades since, detecting new exoplanetary systems has become vital to understanding the nature and variability of other worlds around distant stars. Characterizing the population of exoplanets is key to understanding the mechanisms which drive the formation and evolution of planetary systems both inside and beyond our solar system, and even to understanding the conditions that may allow life to originate on planetary bodies.

The first step in this endeavor is to observe the stars in our local galaxy, and efficiently determine which host exoplanets. Though several methods exist to accomplish this (for instance, measuring the tiny motions of stars due to the gravitational influence of their planets, or directly imaging the planets in the limited cases when this is possible), the method that has so far produced the most detections is called transit photometry. In this method, astrophysicists measure the flux (amount of light) received from a star over a period of time, and attempt to identify the periodic dips in starlight that signify the presence of a planet orbiting between it's host star and our telescopes at Earth.

Never before has the detection and characterization of exoplanets via transit photometry been as promising and feasible as it is now, due to the increasing breadth and sensitivity of time domain optical surveys. Visually identifying transits in stellar lightcurves (flux as a function of time) is impractically time-consuming and tedious, but machine learning is uniquely suited to the task of identifying which lightcurves contain transits. In this project, we propose to explore and evaluate several machine learning algorithms to classify stellar lightcurves as containing or not containing exoplanet transits.

This project is also of personal interest our group because of our fascination both with probing the open questions of astrophysics and with exploring the power and applicability of machine learning to address scientific problems. Thus, we choose this project because it will allow us to learn about the forefront of exoplanet detection science while investigating the world of machine learning.

2 Literature Review

For decades, transits were identified manually in source light-curves with tedious visual inspection (e.g. Charbonneau et al., 2000), which is slow and labor-intensive. The earliest machine-aided detection methods included Box-Fitting-Least-Squares (BLS) algorithms, which scan curves for box-like signals (e.g. Kovács et al., 2002; Grziwa et al., 2012), and Bayesian-based analysis to characterize the likelihood of a signal representing a transit (Aigrain and Favata, 2002). In recent years, interest in supervised machine-learning techniques has risen. These methods usually rely on the previous bodies of human-labeled (sometimes machine-aided) light curves for the generation of training data. Once trained, they can scan through hundreds of curves and flag promising sources for later visual inspection, dramatically reducing the amount of human labor required.

Past work has explored many different types of supervised machine-learning detection techniques, and varying methods of pre-processing input data. Most-commonly, pure time-series light curves (flux measurements recorded over a series of timestamps), are input to an algorithm as features describing an observation.

Other features can be derived from processing the light curves into the frequency domain, for example with Fourier transforms, Wavelet transforms, or phase-folding (see e.g. Stumpe et al., 2014; Pearson et al., 2018). Often, simple (non-ML) algorithms are used to first "triage" curves, flagging transit-like signals for further inspection. Machine learning methods are then applied in a "vetting" phase to predict whether these signals are true transits. Many previously explored algorithms for this task are based on Decision Tree Classification, using simple Decision Trees (e.g. Coughlin et al., 2016; Catanzarite, 2015), Random Forests (e.g. Armstrong et al., 2015; McCauliff et al., 2015), or Gradient Boosted Trees (e.g. Malik et al., 2021). Support Vector Machines and K-Nearest-Neighbors Algorithms have also been implemented with good results (e.g. Schanche et al., 2018). All of these algorithms can achieve good accuracy.

While these basic classification algorithms do perform well, further work has shown that more sophisticated techniques, such as deep learning methods, may achieve even better results in a more streamlined way. Early deep learning models focused on improving previous "vetter" models, starting with a convolutional neural network developed by Shallue and Vanderburg (2018). Other work (e.g. Ansdell et al., 2018; Yu et al., 2019) applied small modifications to this model to incorporate more domain knowledge. More recently, architectures have been developed to detect likely transits without previous triaging. Two one-dimensional convolutional neural network architectures were developed concurrently by Zucker and Giryes (2018) and Pearson et al. (2018). The latter uses phase-folded signals and has been shown to achieve better accuracy on simulated light curves than benchmark BLS and Support-Vector-Machine algorithms (the former was not well-tested against other algorithms). To build on the accuracy achieved with phase-folding, while addressing the issue that it can be difficult to accurately measure the period of a suspected transit, Chintarungruangchai and Jiang (2019) proposed a two-dimensional convolutional neural network that takes as input a 2D stack of all segments (cycles), as opposed to a single averaged phase-folded curve. This model is able to achieve good accuracy even when predicted transit periods are significantly inaccurate.

3 Proposed Work

3.1 Data and Pre-processing

We will use stellar lightcurves from the Kepler mission (Ricker et al., 2015). We chose to use data from the Kepler mission because it is one of the largest exoplanet surveys to date, producing 2708 confirmed detections of transiting exoplanet systems[1]. The program ran from 2009 to 2013, observing approximately 150,000 stars in multiple 90-day quarters, at a cadence of either 30 or 60 seconds between observations. It also used a uniquely high exposure time, and was thus able to observe dimmer, farther away targets than other missions such as the Transiting Exoplanet Survey Satellite (TESS) mission, the other largest exoplanet survey. The mission prioritized Main Sequence stars for which Earth-like planets would be detectable (Batalha et al., 2010). Lightcurves can be publicly downloaded from the Barbara A. Mikulski Archive for Space Telescopes (MAST) archive (DOI: 10.17909/T9059R). The Kepler Science Processing Pipeline is described in Jenkins et al. (2010). Lightcurves comprise measured flux as a function of time. They often contain extended intervals of missing data (due to the telescope entering safe mode, rotating towards Earth, or executing a quarterly roll) as well as individual data points flagged for quality issues (due to cosmic ray hits, reaction wheel zero crossings, impulsive outliers, thruster firings, etc.) (Thompson et al., 2016) which will be one challenge for this project.

Kepler lightcurves are publicly available online through the MAST database, and can be accessed using the interface provided by the `LightKurve` package in Python. Kepler's observing run is divided into quarters, punctuated by rolls of the telescope. We will fit and remove a linear trend from each quarter, and stitch the quarters into one continuous lightcurve. We will then mask out data with quality issues flagged during Kepler data acquisition.

We will also use the publicly available NASA Exoplanet Archive database [2] to create a labeled training dataset. We will cross-reference the target names with the Kepler ID's in this database to label lightcurves in our training set as "confirmed" positive observations (ie. visual inspection or follow-up observing has confirmed the presence of a transiting exoplanet in the stellar system), or "false positive" negative observations

[1] As of June 6, 2022, as reported by the NASA Exoplanet Archive (https://exoplanetarchive.ipac.caltech.edu)
[2] https://catcopy.ipac.caltech.edu/dois/doi.php?id=10.26133/NEA4

(ie. visual inspection or follow-up observing has shown that the lightcurve does not contain transits). We will have to be careful about class imbalance, because many more lightcurves do not contain transits than do.

3.2 Knowledge Extraction

Our task is essentially one of classification. We plan to classify our lightcurves by whether or not they are likely to contain exoplanet transits. Instead of performing a "vetting" on pre-"triaged" curves (see section 3.2), we will work with a full set of Kepler observations. Extracting accurate transit predictions is key to future studies of population-wide exoplanet system statistics, and for follow-up studies of individual systems and planets.

Another interesting knowledge extraction task to which our dataset would be amenable to is an anomaly search. Lightcurves contain a wealth of information not only about the presence or absence of exoplanets, but also about the star itself. Unusual lightcurves have led to breakthrough studies of processes like stellar flaring in the past, or have even been proposed as a potential technosignature in the Search for Extraterrestrial Intelligence (SETI) (e.g. Kipping and Teachey, 2016; Arnold, 2005). Thus, a possible extension of our project would be to serach for anomalous lightcurves.

3.3 Methodology

The wide-range of previously explored algorithms for transit detection makes us excited to cast our net wide as well. First, we will test traditional machine learning algorithms such as K-Nearest-Neighbors, Random Forests, Logistic Regression, and Support Vector Machines. We will apply various transformations to our data before applying these algorithms, namely perform phase-folding and construction of periodograms. We will test algorithms trained on raw, phase-folded, and periodogram data, and present various evaluation metrics (see Section 4) for each model and training set.

In addition to traditional supervised algorithms, we will develop a convolutional neural network model using the PyTorch deep learning framework. The model architecture will be based on that of Pearson et al. (2018). Ideally, after first constructing a baseline model taking as input the pure light curve data, we will (following Pearson et al., 2018) also construct one that intakes Wavelet-transformed data. If this is successful, a model with multiple input channels (pure and transformed light curves) may be attempted.

4 Evaluation

Being a classification problem we would primarily be interested in the following evaluation metrics:

- **Accuracy**: Accuracy is the most straightforward metric and measures the overall correctness of the predictions by calculating the ratio of correctly classified instances to the total number of instances.
- **Precision**: Precision calculates the ratio of true positives (correctly predicted positive instances) to the sum of true positives and false positives (incorrectly predicted positive instances). It indicates how many of the predicted positive instances are actually relevant. High precision indicates that the model is correctly identifying exoplanets and minimizing false positives.
- **Recall** (Sensitivity or True Positive Rate): Recall calculates the ratio of true positives to the sum of true positives and false negatives (missed positive instances). It indicates the proportion of actual positive instances that are correctly identified. High recall indicates that the model effectively captures exoplanets and minimizes false negatives.
- **F1 Score**: The F1 score is the harmonic mean of precision and recall. It provides a single metric that balances both precision and recall. The F1 score is useful to find an optimal balance between precision and recall.

The accuracy metric is inadequate for evaluating the performance of a planet detection algorithm due to the imbalanced nature of most exoplanet detection datasets. These datasets typically have a larger number of light curves without any planet signal compared to those with a planet. While high precision ensures

that the predicted "planet candidates" are mostly true, it is not a very informative metric. This is because achieving high precision can be done by making only a few "planet candidate" predictions and ensuring their correctness. However, this approach may result in missing many potential planet candidates.

Instead, prioritizing recall is more important for assessing the algorithm's performance in planet detection. Recall measures the proportion of actual planet signals that are correctly identified by the algorithm. In our case, it is preferable to accept a higher number of false positives (incorrectly identified planet candidates) rather than missing potential planet signals.

The trade-off between precision and recall, commonly known as the precision-recall trade-off, is significant. A model with high recall may have a lower precision and vice versa. This trade-off is typically evaluated using the F1 score, which combines precision and recall in a single metric.

5 Project Planning

5.1 Collaboration

Our team consists of four members, and as a whole we are well suited to this project. Anna Zuckerman is an astrophysics PhD student with a background in exoplanet photometry, and thus will be well-placed to provide domain knowledge and an understanding of the opportunities provided by the data, as well as its complexities and limitations. She will also guide the knowledge mining goals of the project. Ashutosh Gandhi is a computer science graduate student with both academic and industry experience in machine learning, which will allow him to play a key role in the technical and software development aspects of the project. Andrew Floyd has a computer science and engineering background, with wide ranging experiences and interests. His experience in many types of data mining will be important to developing the methodology for our project and evaluating our models and results. Leah Zuckerman is an astrophysics PhD student, and though her research focus has never included exoplanets she has the unique experience of applying machine learning algorithms to various astrophysical problems. Thus she will play a key role in integrating the scientific and technical aspects of this project.

This project will also benefit from advice and networking with experts in the field. Leah Zuckerman works closely with a Post-Doctoral researcher in Machine Learning at the National Solar Observatory, who will be aptly placed to provide guidance on machine learning best practices. Leah and Anna Zuckerman also have a strong network of peers in the Astrophysics PhD program who are experts in explanatory science. These students may provide feedback on the scientific validity of our methods and results.

5.2 Milestones and Timeline

We will make sure to start early in the semester to allow ourselves time to get feedback and to let the project evolve as we develop our methodology. We will be sure to communicate as a team (we will use the platform Discord to facilitate streamlined and efficient communication) in order to collaborate effectively. We know from past experience how important it is to work effectively as a group.
Our proposed timeline of work is as follows:

- 6/26/23 • Submit this proposal
- 6/30/23 • Finished data exploration and investigation of various feature engineering methods
- 7/03/23 • Finished research of different ML techniques and basic implementation of each
- 7/14/23 • Begun evaluation of each potential ML algorithm + optimal feature engineering
- 7/14/23 • Submit Project Progress Report
- 7/20/23 • Finalize evaluation of each algorithm and determine which is most effective
- 7/21/23 • Draft of the final report and presentation slides
- 7/24/23 • Submit Project Final Report and Presentation

6 Conclusion

The groundbreaking detection of the exoplanetary system PSR1257 + 12 in 1992 marked a significant milestone in understanding planets outside our solar system. Since then, the detection of new exoplanetary

systems has become crucial for studying the formation and evolution of planetary systems and investigating conditions for life. Transit photometry, measuring periodic dips in starlight, has been the most successful method for exoplanet detection. Recent advancements in time-domain optical surveys have made transit photometry even more promising. In this project, the group aims to explore machine learning algorithms to automate the identification of exoplanet transits in stellar lightcurves, combining the frontiers of exoplanet science and the power of machine learning.

References

S. Aigrain and F. Favata. Bayesian detection of planetary transits. A modified version of the Gregory-Loredo method for Bayesian periodic signal detection. , 395:625–636, Nov. 2002. doi: 10.1051/0004-6361:20021290.

M. Ansdell, Y. Ioannou, H. P. Osborn, M. Sasdelli, J. C. Smith, D. Caldwell, J. M. Jenkins, C. Räissi, D. Angerhausen, and and. Scientific domain knowledge improves exoplanet transit classification with deep learning. *The Astrophysical Journal*, 869(1):L7, dec 2018. doi: 10.3847/2041-8213/aaf23b.

D. J. Armstrong, J. Kirk, K. W. F. Lam, J. McCormac, H. P. Osborn, J. Spake, S. Walker, D. J. A. Brown, M. H. Kristiansen, D. Pollacco, R. West, and P. J. Wheatley. K2 variable catalogue – II. machine learning classification of variable stars and eclipsing binaries in k2 fields 0–4. *Monthly Notices of the Royal Astronomical Society*, 456(2):2260–2272, dec 2015. doi: 10.1093/mnras/stv2836.

L. F. A. Arnold. Transit Light-Curve Signatures of Artificial Objects. , 627(1):534–539, July 2005. doi: 10.1086/430437.

N. M. Batalha, W. J. Borucki, D. G. Koch, S. T. Bryson, M. Haas, T. M. Brown, D. A. Caldwell, J. R. Hall, R. L. Gilliland, D. W. Latham, S. Meibom, and D. G. Monet. SELECTION, PRIORITIZATION, AND CHARACTERISTICS OF iKEPLER/i TARGET STARS. *The Astrophysical Journal*, 713(2):L109–L114, mar 2010. doi: 10.1088/2041-8205/713/2/l109.

J. Catanzarite. Autovetter Planet Candidate Catalog for Q1-Q17 Data Release 24. *Astronomy and Astrophysics*, July 2015.

D. Charbonneau, T. M. Brown, D. W. Latham, and M. Mayor. Detection of Planetary Transits Across a Sun-like Star. , 529(1):L45–L48, Jan. 2000. doi: 10.1086/312457.

P. Chintarungruangchai and I.-G. Jiang. Detecting exoplanet transits through machine-learning techniques with convolutional neural networks. *Publications of the Astronomical Society of the Pacific*, 131(1000): 064502, may 2019. doi: 10.1088/1538-3873/ab13d3.

J. L. Coughlin, F. Mullally, S. E. Thompson, J. F. Rowe, C. J. Burke, D. W. Latham, N. M. Batalha, A. Ofir, B. L. Quarles, C. E. Henze, A. Wolfgang, D. A. Caldwell, S. T. Bryson, A. Shporer, J. Catanzarite, R. Akeson, T. Barclay, W. J. Borucki, T. S. Boyajian, J. R. Campbell, J. L. Christiansen, F. R. Girouard, M. R. Haas, S. B. Howell, D. Huber, J. M. Jenkins, J. Li, A. Patil-Sabale, E. V. Quintana, S. Ramirez, S. Seader, J. C. Smith, P. Tenenbaum, J. D. Twicken, and K. A. Zamudio. Planetary Candidates Observed by Kepler. VII. The First Fully Uniform Catalog Based on the Entire 48-month Data Set (Q1-Q17 DR24). , 224(1):12, May 2016. doi: 10.3847/0067-0049/224/1/12.

S. Grziwa, M. Pätzold, and L. Carone. The needle in the haystack: searching for transiting extrasolar planets in CoRoT stellar light curves. , 420(2):1045–1052, Feb. 2012. doi: 10.1111/j.1365-2966.2011.19970.x.

J. M. Jenkins, D. A. Caldwell, H. Chandrasekaran, J. D. Twicken, S. T. Bryson, E. V. Quintana, B. D. Clarke, J. Li, C. Allen, P. Tenenbaum, H. Wu, T. C. Klaus, C. K. Middour, M. T. Cote, S. McCauliff, F. R. Girouard, J. P. Gunter, B. Wohler, J. Sommers, J. R. Hall, A. K. Uddin, M. S. Wu, P. A. Bhavsar, J. Van Cleve, D. L. Pletcher, J. A. Dotson, M. R. Haas, R. L. Gilliland, D. G. Koch, and W. J. Borucki. Overview of the Kepler Science Processing Pipeline. , 713(2):L87–L91, Apr. 2010. doi: 10.1088/2041-8205/713/2/L87.

D. M. Kipping and A. Teachey. A cloaking device for transiting planets. , 459:1233–1241, June 2016. doi: 10.1093/mnras/stw672.

G. Kovács, S. Zucker, and T. Mazeh. A box-fitting algorithm in the search for periodic transits. , 391: 369–377, Aug. 2002. doi: 10.1051/0004-6361:20020802.

A. Malik, B. P. Moster, and C. Obermeier. Exoplanet detection using machine learning. *Monthly Notices of the Royal Astronomical Society*, dec 2021. doi: 10.1093/mnras/stab3692.

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 OFiKEPLER/iPLANETARY TRANSIT CANDIDATES. *The Astrophysical Journal*, 806(1):6, jun 2015. doi: 10.1088/0004-637x/806/1/6.

K. A. Pearson, L. Palafox, and C. A. Griffith. Searching for exoplanets using artificial intelligence. , 474(1): 478–491, Feb. 2018. doi: 10.1093/mnras/stx2761.

G. R. Ricker, J. N. Winn, R. Vanderspek, D. W. Latham, G. Á. Bakos, J. L. Bean, Z. K. Berta-Thompson, T. M. Brown, L. Buchhave, N. R. Butler, R. P. Butler, W. J. Chaplin, D. Charbonneau, J. Christensen-Dalsgaard, M. Clampin, D. Deming, J. Doty, N. De Lee, C. Dressing, E. W. Dunham, M. Endl, F. Fressin, J. Ge, T. Henning, M. J. Holman, A. W. Howard, S. Ida, J. M. Jenkins, G. Jernigan, J. A. Johnson, L. Kaltenegger, N. Kawai, H. Kjeldsen, G. Laughlin, A. M. Levine, D. Lin, J. J. Lissauer, P. MacQueen, G. Marcy, P. R. McCullough, T. D. Morton, N. Narita, M. Paegert, E. Palle, F. Pepe, J. Pepper, A. Quirrenbach, S. A. Rinehart, D. Sasselov, B. Sato, S. Seager, A. Sozzetti, K. G. Stassun, P. Sullivan, A. Szentgyorgyi, G. Torres, S. Udry, and J. Villasenor. Transiting Exoplanet Survey Satellite (TESS). *Journal of Astronomical Telescopes, Instruments, and Systems*, 1:014003, Jan. 2015. doi: 10.1117/1.JATIS.1.1.014003.

N. Schanche, A. C. Cameron, G. Hé brard, L. Nielsen, A. H. M. J. Triaud, J. M. Almenara, K. A. Alsubai, D. R. Anderson, D. J. Armstrong, S. C. C. Barros, F. Bouchy, P. Boumis, D. J. A. Brown, F. Faedi, K. Hay, L. Hebb, F. Kiefer, L. Mancini, P. F. L. Maxted, E. Palle, D. L. Pollacco, D. Queloz, B. Smalley, S. Udry, R. West, and P. J. Wheatley. Machine-learning approaches to exoplanet transit detection and candidate validation in wide-field ground-based surveys. *Monthly Notices of the Royal Astronomical Society*, 483(4): 5534–5547, nov 2018. doi: 10.1093/mnras/sty3146.

C. J. Shallue and A. Vanderburg. 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):94, jan 2018. doi: 10.3847/1538-3881/aa9e09.

M. C. Stumpe, J. C. Smith, J. H. Catanzarite, J. E. Van Cleve, J. M. Jenkins, J. D. Twicken, and F. R. Girouard. Multiscale Systematic Error Correction via Wavelet-Based Bandsplitting in Kepler Data. , 126 (935):100, Jan. 2014. doi: 10.1086/674989.

S. Thompson, D. Fraquelli, J. E. van Cleve, and D. A. Caldwell. Kepler: A search for terrestrial planets (kepler archive manual). may 2016.

A. Wolszczan and D. A. Frail. A planetary system around the millisecond pulsar PSR1257 + 12. , 355 (6356):145–147, Jan. 1992. doi: 10.1038/355145a0.

L. Yu, A. Vanderburg, C. Huang, C. J. Shallue, I. J. M. Crossfield, B. S. Gaudi, T. Daylan, A. Dattilo, D. J. Armstrong, G. R. Ricker, R. K. Vanderspek, D. W. Latham, S. Seager, J. Dittmann, J. P. Doty, A. Glidden, and S. N. Quinn. Identifying exoplanets with deep learning. III. automated triage and vetting of iTESS/i candidates. *The Astronomical Journal*, 158(1):25, jun 2019. doi: 10.3847/1538-3881/ab21d6.

S. Zucker and R. Giryes. Shallow Transits—Deep Learning. I. Feasibility Study of Deep Learning to Detect Periodic Transits of Exoplanets. , 155(4):147, Apr. 2018. doi: 10.3847/1538-3881/aaae05.