Beyond our Solar System: Identifying Exoplanets using Machine Learning

Ryan Faulkenberry Faheem Dustin Quazi, Pratheek Avula, and Ankur Shah
Electrical and Computer Engineering Department, University of Houston, Houston, TX, USA

***Abstract*—Exoplanet Detection is a key field to understanding our universe, and whether we are the only ones in the universe. The primary goal of exoplanet detection is to determine if a star has an orbiting exoplanet, based on time-series data acquired about the star from telescopes or other sensor arrays. This paper presents a general overview of the most common processes used for exoplanet detection, and provides a literature review of the most popular papers which apply machine learning techniques to automate the processes. In addition, this paper presents a potential alternative to existing works, by training and testing a non-pre-trained transformer model on time-series Kepler flux data. The results indicate transformer networks could be a promising way forward for exoplanet detection with time-series data, however the network is outperformed by other literature, likely due to the lack of pre-training and positive training data in the dataset used.**

***Index Terms*—exoplanet detection, machine learning, transformer network, literature review**

I. Introduction

Mankind has looked up to the stars for centuries. It was how humanity was able to established the heliocentric model for our solar system, and ultimately has provided inspiration for works of art and allowed us to better understand the universe we live in. One of the largest questions which exist in modern society is ”are we alone in the universe?”. While there have been many theories, a popular and scientifically rigorous method to attempt to find more planets like our own is known as exoplanet detection and analysis. An exoplanet is any planet beyond our solar system, including rogue (free-floating) planets, and those orbiting other stars. The observable universe is estimated to be around 93 billion light years in diameter [1], and the Milky Way alone contains an estimated 100 billion stars. The following big question can be extracted in the quest for searching for life: Among these stars, how can we detect the presence of exoplanets around stars in order to narrow down our search for life?

This paper intends to provide a general overview of exoplanet detection methods and document current literature covering how machine learning has been applied to exoplanet detection. Based on the literature review, the authors also made an attempt at developing models based on

The paper is structured as follows: Section II provides an overview of traditional methods used to detect exoplanets. Section III provides an overview and general literature review of how ML has been applied to exoplanet detection, and the processes used. Section IV discusses the Transformer model used for the experiment in this paper and the approach taken for our Exoplanet Detection system. Section V provides the results to the model defined in the previous section and findings as to how the models selected fared compared to literature. Finally, Section VI draws general conclusions to the work and provides some potential future directions for the work.

II. How Exoplanet Detection Works

There are a few methods used to detect exoplanets. Almost all methods use some form of imagery of a star in order to draw different conclusions about the planet. In all these cases, a ”pass” in the test indicates the presence of an exoplanet, and typically one additional piece of information. The following section describes the most popular methods used.

The radial velocity method identifies the movement of a star in response to the tug of a planetary body. This is accomplished by observing a change in the position of the star using spectral analysis [3], such as with Doppler shifting [4]. This is visualized in Figure 1. This was the most popular method used to detect exoplanets prior to the development of the Transit Method. It is limited due to its reliance on the axis of ”wobble” being in line with the observer, and that if a star has multiple exoplanets, this method alone would not be able to establish the difference between one or many exoplanets [5].

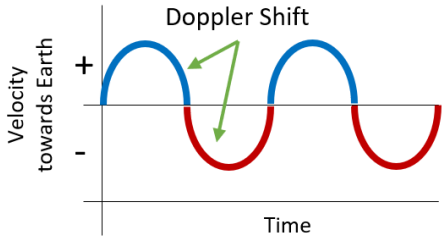

Fig. 1. Visualization the Radial Velocity Method on a plot. The color represents Doppler shifting as a result of the star moving towards or away from Earth, indicating the presence of a mass affecting the position of the star of interest.

The transit method uses imagery of a star taken over time to identify a significant, periodic, dip in the magnitude of the brightness of the star. Depending on the size and intensity of the dip, the mass of a planet can be determined. Missions such as the NASA Kepler Space Telescope were designed specifically to use this method in order to detect exoplanets [2], and the data it produces can be plotted as shown in Figure 2. While this method is better at identifying multiple exoplanets, it does require a much higher resolution telescope or sensor

array to be able to detect minute changes in brightess at stellar distances.

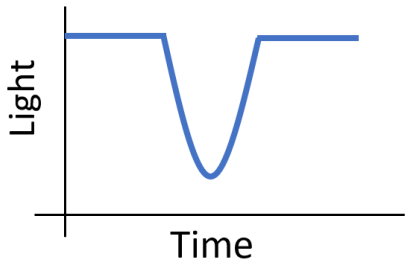

Fig. 2. Visualizing the Transit Method on a plot. The dip in brightness (light value) represents a body passing between the telescope and the star of interest.

The direct imaging method involves directly taking a picture of the planet [3]. This is a relatively new method, and relies on ”squelching” [3] the light of a star in order to directly see exoplanets in the vicinity. To date, only 69 planets have been identified using the direct imaging method [6].

III. Literature: ML for Exoplanet Detection

In all three methods of detection, time-series imagery of the sky (or data acquired from multi-sensory telescopes that is then presented as false-color imagery) are used to draw conclusions about the existence of exoplanets. Such data is ripe to be analyzed with Machine Learning techniques, however there are very few confirmed positive cases relative to the total data size: There are around 5,500 confirmed exoplanets in the Exoplanet Archive [6], inclusive of all experiments and telescopes being used for this purpose, however in just the data set many works use (Kepler Space Telescope experiment), there are approximately 200,000 stars [7]. With this limitation, some works reviewed [12] [13] used the Synthetic Minor Oversampling (SMOTE) data augmentation technique to address the imbalance in class distribution, namely to randomly replicate the confirmed exoplanet cases.

Google Scholar states there are approximately 6,600 papers when searching for ”exoplanet detection machine learning”. The state-of-the-art method that many works cite, authored by Shallue and Vanderburg, uses a convolutional deep neural network trained on light curves similar to those used by our work [8]. This work tested a linear (zero hidden layers), fully connected, and convolutional architecture, and found that the convolutional architecture worked best [11].

Another work which used this data-set worked with the LightGBM Gradient Boosted Trees model - unlike the previous work [11], this work was trained on three separate measurements of the same star, as well as simulated data. LightGBM is a model which uses the weighted average of a chain of decision trees to produce a binary output [14]. While the overall accuracy and AUC was lower than the state-of-the-art [**?**], it had a better accuracy than both their Linear and Fully Connected architectures, and a comparable AUC score to their linear model. This demonstrates that there may be potentially cheaper ways to accomplish reasonably accurate exoplanet identification.

A similar work used Majority Voting technique of Ensemble Learning for the task in order to try and obtain more precise classifications. After testing multiple network types, the authors used the highest performing networks (SVC, Random Forest, KNN, and MLP) and used majority voting, where the class label with the greatest frequency from all models is the final output [12]. This work demonstrated that traditional (non-deep-learning) ML methods may be feasible for exoplanet detection, especially considering that every model they tested had accuracy scores higher than 90 percent, and the top performers were above 97 percent.

Other works deviated slightly from just identification, and went into habitability (answering the big picture question). This is accomplished by using additional sensor readings extracted from telescope imagery or by deriving parameters (size, mass, distance from star) which are a function of the sensor reading. While these works generally are applied to confirmed exoplanets, we include these works as their approach may be beneficial for the detection step. One work in particular used a variable auto-encoder with anomaly detection, where anomalous results indicated the potential for an exoplanet to be habitable [13]. Finally, another work proposes a deep-learning model called ASTRONET which combines work done in the aformentioned works to create a pipeline for both detection and habitability, using fully connected networks [15].

IV. Our Attempt: Transformer Network

Similar to other exoplanet detection literature, for this experiment, a dataset consisting of labeled observations from the Kepler Space Telescope was used. This dataset [8] consists of flux values (measurements of brightness) over one full observational period of the Kepler experiment, pre-processed and normalized for noise reduction. This data is highly imbalanced, with under 1% of the set containing positive cases (that is, cases in which there is a confirmed exoplanet).

To account for the imbalanced dataset, we use the Synthetic Minority Oversampling Technique (SMOTE) [9] to generate synthetic positive cases. After SMOTE, our training data is nearly double the original size, and the classes are nearly perfectly balanced.

To classify this time-series data in a novel way, we are motivated by the successful application of the transformer to other kinds of sequential data. We design a transformer model tailored to time-series data and call it “Keplerformer.”

To understand the function of *Keplerformer*, consider its parent model, Vision Transformer (ViT) [10]. ViT operates on images by splitting them into patches and projecting each patch to a higher dimension. Upon this set of projected patches, Multi-head self-attention is applied, in which multiple attention heads build separate representations of the patch sequence via self-attention. One head may model relationships between nearby patches, while another head may model relationships between distant patches for example. This process is repeated multiple times, then the output of the final attention

block is passed to a multi-layer perceptron to provide classifier output.

Keplerformer simply adapts this model to time-series data. One entry from the Kepler dataset is about 1700 flux points collected linearly through time. Whereas ViT would form patches from an image, we form patches from each group of 4 adjacent flux readings in the time series data. Linear projection to a higher dimension is skipped, because the number of points per patch is small, and the data were collected independently.

V. Experiment and Results

We train *Keplerformer* for 100 epochs using the default parameters from ViT. *Keplerformer* applied on the Kepler dataset performs with 93.3% validation accuracy. However, as is shown in Figure 3, our model failed to correctly classify any of the 5 cases in which an exoplanet was present. This is unfortunate but not surprising: it is known that transformers tend to be very dependent on pre-training, and training in unusual domains can be a long and arduous process. It is worth noting here that fewer than 1% of the data was for the positive case; had the model had additional cases to classify, it may have performed better on positive cases.

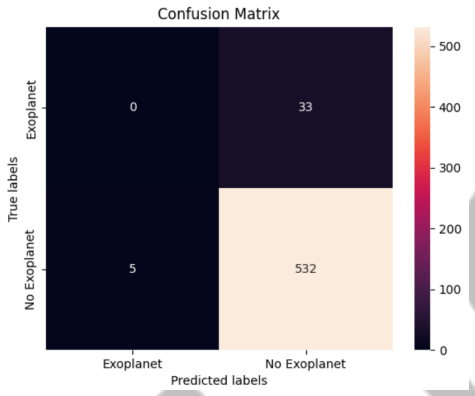

Fig. 3. Confusion matrix demonstrating classwise predictions. The model is conservative: it is good at classifying negative cases, but struggles to classify positive cases.

VI. Conclusion

In this work we described the task of exoplanet detection and the methods applied to this task, both classical and recent. The transit method was the most popular method used by ML literature, and the most common in terms of confirmed exoplanet discoveries. Popular works would use both traditional and deep-learning machine learning methods in order to accomplish detection, but convolutional neural networks appeared to be the most popular.

Our work applied a transformer network to a pre-processed dataset consisting of time-series star flux values from the Kepler Space Telescope experiment; while it correctly classified most negative cases, it was unable to classify any of the few positive cases.

There is room for future work to improve out model. For example, pre-training the transformer with self-supervised learning on the huge volume of Kepler data before supervised training could help the model learn the features of the data necessary to make good classifications.

References

[1] NASA, "What is an Exoplanet?," Exoplanet Exploration: Planets Beyond Our Solar System, Apr. 02, 2021. https://exoplanets.nasa.gov/what-is-an-exoplanet/overview/

[2] C. S. Agency, "Detecting exoplanets," Canadian Space Agency, Apr. 22, 2022. https://www.asc-csa.gc.ca/eng/astronomy/beyond-our-solar-system/detecting-exoplanets.asp

[3] NASA, "How We Find and Characterize — Discovery," Exoplanet Exploration: Planets Beyond our Solar System, Apr. 13, 2022. https://exoplanets.nasa.gov/discovery/how-we-find-and-characterize/

[4] L. Lindegren and D. Dravins, "The fundamental definition of 'radial velocity,'" Astronomy and Astrophysics, vol. 401, no. 3, pp. 1185–1201, Apr. 2003, doi: https://doi.org/10.1051/0004-6361:20030181.

[5] "Color-Shifting Stars: The Radial-Velocity Method," The Planetary Society, 2020. https://www.planetary.org/articles/color-shifting-stars-the-radial-velocity-method

[6] "Exoplanet and Candidate Statitics," California Insitude of Technology / NASA, 2019. https://exoplanetarchive.ipac.caltech.edu/docs/counts_detail.html

[7] "Kepler," MAST, https://archive.stsci.edu/missions-and-data/kepler (accessed Dec. 4, 2023).

[8] "Exoplanet Hunting in Deep Space," www.kaggle.com. https://www.kaggle.com/datasets/keplersmachines/kepler-labelled-time-series-data

[9] Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. 2002, 16, 321–357.

[10] Dosovitskiy et. al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," International Conference on Learning Representations, dio: https://doi.org/10.48550/arXiv.2010.11929

[11] 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, vol. 155, no. 2, p. 94, Jan. 2018, doi: https://doi.org/10.3847/1538-3881/aa9e09.

[12] G Rakesh, M. Jahnavi Bhuvana Chandrika, Ch., and S. Manne, "Exoplanet Detection Using Feature Engineering with Ensemble Learning," Jun. 2023, doi: https://doi.org/10.1109/icpcsn58827.2023.00025.

[13] Y. Patel, S. Tiwari, Sanjay Kumar Sonbhadra, and S. Agarwal, "Predicting Habitable Exoplanets in Different Star-Systems Using Deep Learning Based Anomaly Detection Approach," Jun. 2023, doi: https://doi.org/10.1109/ijcnn54540.2023.10191791.

[14] A. Malik, B. P. Moster, and C. Obermeier, "Exoplanet detection using machine learning," Monthly Notices of the Royal Astronomical Society, Dec. 2021, doi: https://doi.org/10.1093/mnras/stab3692.

[15] R. Jagtap, U. Inamdar, S. Dere, M. Fatima, and N. B. Shardoor, "Habitability of Exoplanets using Deep Learning," 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Apr. 2021, doi: https://doi.org/10.1109/iemtronics52119.2021.9422571.