The Wayback Machine - https://web.archive.org/web/20210312033453/https://csfjournal.com/volume-3-issue-5…

*February 28, 2021*

# USING MACHINE LEARNING TO REMOVE NOISE FROM STELLAR SPOTS IN EXOPLANETARY DATA FROM SPACE TELESCOPES

ARTASH NATH

Grade 8 | Toronto, Ontario

**INSPO North America Science Fair 2020 Junior Gold Medal | NASA Space Apps COVID-19 Challenge Global Winner**

*Exoplanets are planets around other stars. When an exoplanet transits in front of its parent star, its main body blocks out some light of the star. This causes a dip in the light received. If the exoplanet has an atmosphere around it, then the atmosphere will also absorb some of this light. How much light is absorbed by the atmosphere depends on the wavelength of the light, and the thickness and gases present in the atmosphere. If we plot the transit of an exoplanet in different wavelengths, we will get light curves of different depths. Studying transit light curve depths of exoplanets in different wavelengths allows us to predict the chemical composition of their atmospheres.*

*The presence of stellar spots adds noise to this data. Stellar spots are cooler and darker than the surrounding surface. They may overlap with path of the transiting exoplanet and corrupt the transit light curves data. It would lead to errors in the calculations of the planet-star radius ratio. We must remove the effects of stellar spots on the transit light curves from those produced by the exoplanetary atmospheres. The current approach is to identify the effects of spots visually and correct for them manually - which is time consuming and prone to errors, or to discard the data.*

*I created a hybrid machine learning model to remove the effects of stellar spots in faint signals of transiting exoplanets' atmospheres received by the space telescopes. My model was able to accurately predict the exoplanet-star radius ratios in 55 wavelengths with a mean square error of 0.001. The model works well in low computational power environments. It can be applied to data from space and ground based telescopes like NASA's Transiting Exoplanet Survey Satellite (TESS) and the upcoming Vera C. Rubin Observatory.*

## INTRODUCTION

Over 4,000 exoplanets have been discovered so far and there may be billions more in our galaxy alone. To answer some of the bigger questions about exoplanets – their formation, the atmospheric conditions, and chemical composition of their atmospheres, we need to analyze the data available

about them. Many telescopes are currently collecting exoplanetary data, such as the NASA's Transiting Exoplanet Survey Satellite (TESS) and the European Space Agency's CHaracterising ExOPlanet Satellite (CHEOPS).

More telescopes mean more opportunities to learn about the universe. But telescopes also generate massive amounts of data. For example, the Square Kilometers Array (SKA) Telescope, once operational will produce exabyte-sized dataset daily. The data from telescopes also come in different formats: images, light curves data, spectroscopic and radio observations.

*While data about our universe is increasing exponentially, the astronomy community is not.* This means custom machine learning models are needed to extract useful information from big datasets. Machine learning algorithm can classify data based on knowledge it has been previously trained on (the training dataset). Once a machine learning algorithm has been accurately trained it can be applied to make predictions in newly generated data from telescopes with no supervision.




Figure 1

## PROBLEM STATEMENT

**Stellar Spots Add Noise to Exoplanetary Data Leading to Incorrect Calculation of the Planet-Star Radius Ratio**

When an exoplanet transits in front of its parent star, it blocks out some light from the star. The

telescope observing the star would register a dip in the incoming light, i.e. the brightness of the star would seem to decrease. If we were to plot the change in incoming light from the star over time, we would end up with a transit light curve. By measuring the depth of the light curve, we can calculate the relative planet-star radius ratio.

If the exoplanet has an atmosphere around it, then the atmosphere will also absorb some of this light. How much light is absorbed by the atmosphere depends on its thickness and the composition of gases present. When light of a given wavelength is absorbed by the exoplanetary atmosphere, the transit light curve will dip more, and the radius of the exoplanet will appear bigger in that wavelength. For wavelengths that are not absorbed by the exoplanet's atmosphere, their transit light curves depth will be shallower, and the radius of the exoplanet will appear smaller. Observing transit depths in multiple wavelengths allows us to get relative planet-star radius ratios in different wavelengths. *See Figure 1.*



Figure 2

Figure 2

The presence of stellar spots on the surface of the star affects its brightness. Stellar spots are cooler and darker than the surrounding surface. Their overlap with transiting exoplanet corrupts the transit light curves and leads to incorrect calculations of the planet-star radius ratio. This noise needs to be removed so that we can separate the dips in light caused by exoplanetary atmospheres from the effects of stellar spots. *See Figure 2.*

But this is a hard problem. The current approach is to identify the effects of spots visually and correct for them manually or discard the data (Nikolaou et al., 2020).

## HYPOTHESIS

Machine learning models can be trained to remove noise from stellar spots in exoplanetary transit light curves. They can then accurately predict the planet-star radius ratios of exoplanets in different wavelengths.

## DATASETS USED

The dataset for the project was provided by the Atmospheric Remote-sensing Infrared Exoplanet Large- survey (ARIEL) Telescope. ARIEL is the European Space Agency's (ESA) first mission dedicated to measuring the chemical composition and thermal structures of hundreds of transiting exoplanets. It will study what exoplanets are made of, how they form and how they evolve, by surveying a diverse sample of about 1000 exoplanets, simultaneously in visible and infrared wavelengths**.**

I used 150,000 simulated exoplanetary observations available on the ARIEL Space Telescope website. For each of the exoplanets, their transit light curves in 300 time-step data-points over 55 different wavelengths were provided. In addition, six stellar and planetary parameters were provided - mass, radius, temperature, log, period, and magnitude of the stars. Included in this database was also the planet-star radius ratios for each of the 55 wavelengths. *See Figure 3.*

The dataset is publicly available and can be accessed at: https://ariel-datachallenge.azurewebsites.net/ML



Figure 3

Figure 3

## PROCEDURE

As the dataset provided by the ARIEL Telescope was already labeled with the planet-star radius ratios for each exoplanet, I could use a supervised machine learning algorithm. In supervised learning, the algorithm is taught using a training dataset what output results it should predict for specific inputs. The trained model then predicts outputs for a new dataset.

I chose the Long Short-Term Memory (LSTM) model - a form of Recurrent Neural Network (RNN), to make the planet-star radius ratio predictions. RNNs, are a class of neural networks that allow outputs from previous steps to be used as inputs to the next step while maintaining hidden states. In a LSTM every subsequent layer encodes deeper and more complex information. LSTM work very well for sequential data as cells in each layer learns what to remember and what to forget. They assign weightage and focus on those parts of the data that are affected by the noise. My LSTM model had 2 LSTM Layers, each containing 256 nodes.

**Step 1: Transforming Exoplanetary Data Into a 3D Array**
I imported the entire dataset into Python and converted it into a 3-dimensional array. ·

- The *first* dimension was the list of exoplanets for which the light curve was available.
- The *second* dimension was the 55 wavelengths for which the light curve was generated.
- The *third* dimension was the 300-time steps for which the light from the star during the transit event was measured and used to create the light curve.

I split this array into training set and testing set in the ratio of 80:20.

**Step 2: Setting Up the Recurrent Neural Networks**
I decided to use the Keras library with the TensorFlow Framework backend to create my RNN Model. These libraries make it easier to design custom Neural Networks and accelerate their training using GPUs.

The inputs to my RNN were the 55 sequences of light curves for every exoplanet passing in front of its parent star taken in different wavelengths. This would result in 55 outputs: *the predicted planet to star radius ratio for each wavelength.*

During its training, the RNN would go over the entire training dataset several times or complete several epochs. Each time, the RNN would go over a single data point consisting of one input and one output. It would then analyze the input and learn why a specific output is assigned to it. After enough training, it would be able to analyze inputs it has not seen before and recommend an output based on that with reasonable accuracy.

I ran my RNN for 40 epochs over a period of 3 hours. It trained on a Nvidia GTX1080, with 32 GB of RAM, and an Intel CPU. In each epoch, the RNN goes through and learns from the entire dataset.

## FINDINGS AND INTERPRETATIONS

The hybrid machine learning model – a combination of Long Short-Term Memory (LSTM) and Feed Forward neural network was able to remove noise from stellar spots and accurately calculate planet to star radius ratio.

The LSTM in the hybrid model handles the time series (or the sequential) data such as transit light curves. The Feed-Forward neural network handles the numerical data such as mass, radius, temperatures, period, and magnitude of the stars. A 'Concatenate Layer' is later applied to merge results from the two machine learning models before passing it through 'Dense Layers' to generate the output (planet-star radius ratio). *See Figure 5.*

To compare how accurate my predicted results were, I plotted the Mean Squared Errors vs. Epoch graph. The hybrid model attained a mean squared error (MSE) of 0.00053 on the training dataset and 0.001 on the test dataset. The MSE trend is smooth and always decreasing implying it is always able to improve itself as it completes more epochs. *See Figure 6.*

The algorithm becomes fairly accurate in its prediction after just 15 epochs This means that my model can learn quickly and can be retrained on other similar datasets in environments with limited computing and processing capabilities.



Figure 5

I have prepared graphs of planet-star radius ratio predicted for 55 different wavelengths for a sample of six exoplanets. *See Figure 7.* The initial data for these 6 exoplanets was corrupted by the noise from the stellar spots. After applying my hybrid machine learning model, I was able to remove the effects of the noise so that only the effects of exoplanetary atmospheres are shown. The graphs show how close the predicted values of planet-star radius ratios are from the ground truth for the 55 different wavelengths.



Figure 6

## CONCLUSION

**Machine Learning Models Can Remove Stellar Noise and Accurately Predict Planet-Star Radius Ratios**

The hybrid machine learning model I created was effective in reducing the noise from stellar spots in the exoplanetary data obtained from space telescopes.

1. The model achieved a Mean Squared Error (MSE) of 0.001 in predicting the planet to star radius ratio.
2. The model works on different data types: Time-Series and Numerical.
3. As it is a hybrid Model, it can be expanded to include new input data types such as images. It even works when some data is missing or becomes available later.
4. The model can be trained on a single GPU machine in less than a day making it accessible to astronomy community worldwide, especially those working in low computing and processing power environments.

## OUTREACH AND COMMUNITY BUILDING

**Free Online Training Module for Students and Researchers on Machine Learning and Space Data**

Data from ground and space telescopes are growing exponentially while the astronomy community is not. The data needs to be analyzed for the exoplanetary research field to advance. Applying machine learning could be one way to find patterns quickly and accurately in the data and solve many of the big challenges faced by astronomers.

I wanted to create a community of researchers around machine learning and space telescopes data and excite younger people to apply their machine learning skills to astronomy.

I prepared a free online training module using Jupyter Notebook (in python). The module provides all the steps and exercises needed for researchers to replicate my results or produce results on their own datasets by applying the hybrid machine learning model.

The Module takes 3 hours to complete and over 50 participants from Mexico, Canada, United States, and India have completed it. On 13 October 2019, I used my module to conduct an onsite Workshop at the MIT Media Lab in Massachusetts, USA for the participants of the Global Bio Summit on applying machine learning to the ARIEL dataset.

The online tutorial is available from my GitHub https://github.com/Artash-N/Ariel-Machine-Learning- Training-Module-for-Exoplanets

## FUTURE APPLICATIONS

**Predicting Chemical Composition of Exoplanetary Atmospheres**

My hybrid machine learning model can be applied to near-real time data received from space and ground based telescopes. I plan to extend my model to include data from other space telescopes, including NASA's Transiting Exoplanet Survey Satellite (TESS) and the James Webb Space Telescope (JWST). This will increase the training dataset available for my machine learning algorithm.

The machine learning model can be extended to solving other problems including predicting chemical composition of transiting exoplanets and detecting chemical biosignatures of life. *See Figure 8.* This would push science, exploration, provide new insights on planetary science beyond our solar system, and motivate other students to apply machine learning in astronomy.

I hope to be able to access computers with more Graphic Processing Units (GPUs) to transfer learnings from this project to solve other problems in the field of astronomy and space exploration



Figure 7

Figure 7

## INSPIRATION

I got inspired to do this project after coming across the website of the European Space Agency's Atmospheric Remote-sensing Infrared Exoplanet Large-survey (ARIEL) Space Telescope. ARIEL Telescope will launch in 2028 and will study what exoplanets are made of observing about 1000 exoplanets in 55 visible and infrared wavelengths.

In May 2019, the ARIEL Space Telescope team launched their Data Challenge Initiative and released a dataset of 150,000 simulated exoplanet observations. The objective was to build a global community for exoplanet data solutions that would help them analyze the data received from the telescope.

As I love space and astronomy and had been developing machine learning algorithms on my own, I decided to participate in the challenge. Availability of a big and free dataset of 150,000 observations online meant I could train and test my models on the dataset provided from home. I worked over

several months in my free time after school to understand the dataset, the science of exoplanets, the methods used by astronomers to detect exoplanets and their atmospheres. I combined all this knowledge to create new machine learning models.

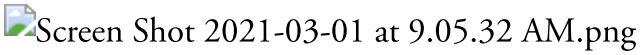
## ACKNOWLEDGEMENTS

I thank Prof. Giovanna Tinetti, Director, University College of London, Centre for Space Exochemistry Data, UK, for inviting me to present the findings of my project at "The ARIEL Science, Mission & Community 2020 Conference." It was held on 14 – 16 January 2020 at the European Space Research and Technology Centre (ESTEC) in the Netherlands. https://www.cosmos.esa.int/web/ariel/conference- 2020

## REFERENCES

Nikolaou, Nikos & Waldmann, Ingo & Sarkar, Subhajit & Tsiaras, Angelos & Morvan, Mario & Yip, Kai & Tinetti, G.. (2020). Correcting Transiting Exoplanet Light Curves for Stellar Spots: A Machine Learning Challenge for the ESA Ariel Space Mission.

## ABOUT THE AUTHOR




**Artash Nath**

I like solving big challenges: from measuring the effectiveness of COVID-19 lockdowns in reducing human movement to determining exoplanetary atmospheres using machine learning. I am passionate about space exploration, rockets, robotics, seismology, and artificial intelligence. In 2020, I won the Junior Gold Medal at the INSPO North America Science Fair 2020, and became the Global Winner of the NASA Space Apps COVID-19 Challenge from amongst 1600 teams.

❤

Newer Post

DragonFly Bot: Learning from Nature using Neurons and Logic Gates (/web/20210312033453/https://csfjournal.com/volume-3-issue-5/2021/2/21/dragonfly-bot-learning-from-nature-using-neurons-and-logic-gates)

Older Post

A Novel Approach for the Bio-Friendly Extraction of Microplastics with Ascidians (/web/20210312033453/https://csfjournal.com/volume-3-issue-5/2021/2/22/a-novel-approach-for-the-bio-friendly-extraction-of-microplastics-with-ascidians)

SUBMIT (/WEB/20210312033453/HTTPS://CSFJOURNAL.COM/TAKE-ACTION)

CONTACT (/WEB/20210312033453/HTTPS://CSFJOURNAL.COM/CONTACT)