Project ID: PHYS-103

Using An AI(GANS) Machine Learning Program to Find Anomalies in Large Astrophysics Data Such As Star Candidates That Could Be SETI Or Other Anomalies

Nicholas Vasilescu, Brooklyn Technical High School

November 17, 2022

1 Introduction

In recent years, astrophysicists and amateur scientists have gained access to large data from satellite telescopes that NASA has sent to space.[1] In particular, NASA has sent two satellites telescopes, Kepler and TESS, to obtain data regarding the amount light over time created by hundreds of thousands of stars in our galaxy. These telescope satellites were created to find exoplanets, planets not in our solar system but in our galaxy. By measuring the amount of light from each star over time, these telescopes can find evidence of planets rotating around various stars. The theory behind this is that if the planets around such stars are on the same plain as our planet and solar system, then the planets will periodically block some of the light from the stars that reach us. By measuring and plotting the light value over time, specifically over 26 days, if the light value periodically reduces in value, then that light curve could evidence an exoplanet rotating around a star. So far, scientists working with NASA have found approximately 5,000 stars that have planets rotating around them and they have published them in a list called TESS Objects of Interest (TOI). Some astrophysicists have theorized that Kepler and TESS data could be used to find evidence of other anomalous stars, such as stars near black holes. Astrophysicist who have examined unusual light curves from stars in our galaxy have found star candidates that could be evidence of nearby alien intelligent civilizations. For decades, scientists have been theorizing about and searching for extraterrestrial intelligence (SETI). In the past decade, using Kepler and TESS data, at least three stars have been found with usual transit light curves that are evidence of possible SETI.

For my research, I set out to see if there is an efficient way to find similar SETI star candidates, or other types of anomalous stars, by searching through TESS data. One of the problems with hunting for such evidence is that the TESS data is massive; covering 50 sectors of the sky, the TESS satellite has already created approximately one million files of data taking up over one terabyte of data storage. For my research, I created and used an artificial intelligence (AI) program to automatically go through the 1 million TESS files to try to find anomaly star candidates, including anomolies that could be evidence of SETI. One version of such AI programs, which are open source and very powerful, are Generative Adversarial Network (GANs) machine learning programs.

My use of a GANs program with astrophysics data appears to be novel. While some astrophysicists have employed artificial neural networks to characterize TESS data concerning if

[1]The most famous new space project by NASA is the James Webb Space Telescope that successfully launched on December 25, 2021 07:20 am EST. The Webb Space Telescope has captured this public's imagination by sending back to Earth massive data and pictures from deep space in the universe. ((NASA, d)). Also capturing the public's imagination on space is NASA's recent testing projectile to destroy or prevent an asteroid hitting Earth is the Dart Project.((NASA, a))

there were additional exoplanets, they have not employed GANs AI programs. See, e.g.: ((Osborn et al., 2020)) ((Ofmana et al., 2021)) ((Friedman, 2020)) ((Tu1 et al., 2022)) ((Ansdell et al., 2018)) Shallue1 and Vanderburg ((2018)). Thus, my use of a GANs program on TESS data might be the first time GANs has been used on astrophysics data.

For my research program, I applied a GANs open source machine program to use on TESS data. The program is in alpha development mode and is a version that MIT graduate students created called Orion-ML. ((to AI Lab at MIT)) Using a GANs model built into Orion-ML called TadGan, I trained my Orion-ML program on the 5,000 TOI files and then used the trained Orion-ML model in a Python program to automatically look through the 1 million files. My program identified files that could be anomalies compared to the TOI files. By creating a model trained on the TOI list of stars, the model scored transits that are anomalous compared to the TOI list. The TadGan model would then find anomalous light curves that are (1) possible anomaly stars that have SETI near by; (2) exceptional exoplanets not already found in the TOI or other transits by stars in our galaxies that are anomalies; or (3) other anomalous stars with strange transit light curves. In terms of evidence of SETI candidates, my program was able to find about 20 candidates that could be examined further to determine if there are SETI nearby, and scores of other anomaly candidates such as large binary exoplanets and/or binary stars.

2 Background

a. Kepler and TESS data

On March 6, 2009, NASA launched the Kepler spacecraft to "watch a patch of space for indications of Earth-sized planets orbiting stars similar to the sun." ((NASA, c)) As NASA explained, the Kepler spacecraft was designed to obtained data of the light value for 150,000 stars over time:

> The area that Kepler will watch contains about 150,000 stars like the sun. Using special detectors similar to those used in digital cameras, Kepler will look for a slight dimming in the stars as planets pass between the stars and Kepler. The observatory's place in space will allow it to watch the same stars constantly throughout its mission, something observatories such as NASA's Hubble Space Telescope and ground-based telescopes cannot do. ((NASA, c))

Kepler retrieved data that evidenced exoplanets by "detecting the tiny brightness dips caused when they pass in front of their stars' faces" from the instrument camera's perspective. ((NASA, c)) The light values plotted over time are called "light curves" and the light curves with repeated dips are called transits. Significantly, Kepler did this work by staying locked onto 150,000-plus target stars using three gyroscope-like devices called reaction wheels. By 2015, having found

thousands of exoplanet candidates, some of Kepler's reaction wheels broke down causing it to lose precision. ((NASA, c))

On April 18, 2016, NASA launched the Transiting Exoplanet Survey Satellite (TESS) to replace Kepler. Using more modern technology, particularly with the digital cameras, TESS was designed to cover an area of space 400 times larger than the area that Kepler covered and stars that were 30 to 100 times brighter than the stars that Kepler studied. The TESS mission broke up the area in space it covers into 26 different sectors, each 24 degrees by 96 degrees across. TESS's cameras stare at each sector for at least 27 days, looking at the brightest stars at a two-minute cadence. By December 21, 2021, the data sent back to Earth by TESS for scientists to study, results in more than 5.000 "objects of interests" ("TOI"), i.e. possible exoplanets in our galaxy. Significantly, the data that NASA obtains from TESS, is made available to the public through a website called The Mikulski Archive for Space Telescopes (MAST) that is maintained in Baltimore, Maryland at the Space Telescope Science Institute (STScI). ((NASA, b)) TESS has expanded and now covers at least 50 sectors of the sky and the data from MAST for the 50 sectors is nearly 2 Terabytes of data and nearly 1 million separate files.

b. Alternate Uses for TESS DATA: Hunting for Black Holes, SETI and other Anomalous Objects.

Some astrophysics scientists have theorized that data from TESS could be used to find stars that are near other objects in the galaxy in addition to stars that have exoplanets. For example, Kento Masuda and Kenta Hotokezaka, two graduate students in the Department of Astrophysical Sciences at Princeton University, theorized in their paper published on September 23, 2019, that TESS data could be used to find star candidates that are near black holes.Masuda and Hotokezaka ((2019)) The paper theorized that stars that are near or captured by black holes would have their light magnified by the gravity by the black hole as the stars rotated around the black hole. According to this paper, the light curves from such stars would not be transit light curves like those for stars with exoplanets, but rather light curves that have repeating pulses in the light curve. Masuda and Hotokezaka ((2019))

One type of stars in our galaxy that TESS data could be used to find is star candidates that are SETI. What is a SETI star? For decades, scientists have theorized that it is highly likely that there are extraterrestrial intelligent civilizations in our galaxy and in the other billions of galaxies and the quest is called Searching for Extraterrestrial Intelligence (SETI). The theoretical physicist and mathematician, Freeman John Dyson (December 15, 1923 – February 28, 2020), postulated that advanced alien intelligent civilizations may create a mega-structure that encompasses a star and captures a large percentage of its solar power. After all, our star

creates more massive energy every second than any contraption on Earth. Scientists seeking evidence of SETI in our galaxy theorize that a Dyson Sphere could change the light over time of the light reaching Earth so the light curves have unusual transits.

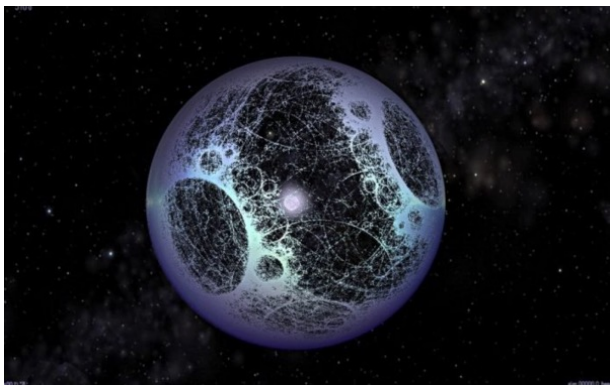


Figure 1: Dyson Sphere rendering – Credit: Developments

More recently, scientists have postulated in a 2015 SETI related article that rather than implement a Dyson Sphere, an alien civilization may illuminate the dark side of a synchronously rotating planet using light from a nearby star. Korpela et al. ((2015)) The article suggests that the mechanism that an alien civilization creates could be the following which could create a certain type of light curve of the nearby sun:

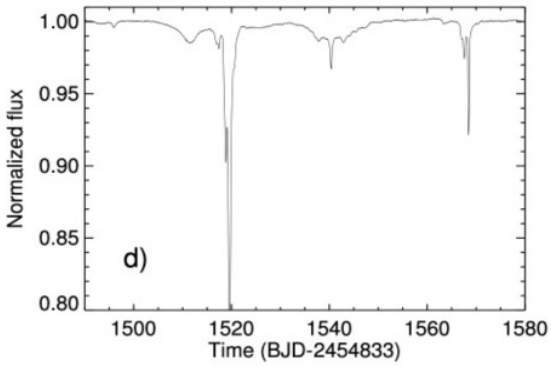
> To provide illumination equivalent to the bright-side stellar illumination, it would need to be a large disk or annulus somewhat larger than the planetary diameter. Since the mirror is locked in the anti-stellar direction it would provide non-uniform directional illumination on the ground, with regions near the terminator receiving illumination at higher incidence angles than regions near the anti-stellar point.((Korpela et al., 2015)).

In 2016, other scientists theorized that an advanced alien civilization could manipulate light from a nearby star with powerful lasers to effect the transit light curve to communicate across the galaxy or universe to communicate with other advance civilizations. See A Transit Signature for SETI? by PAUL GILSTER on APRIL 4, 2016 (a-transit-signature-for-seti).

More recently, starting in 2018, NASA has recognized and starting subsidizing a new research field called "Technosignatures", where scientists look closely at the exoplanets found by the Kepler and Tess satellites to see if there is evidence of alien civilizations.((Kaufman)). The powerful Webb telescope will allow such "Technosigatures" scientists to now look very closely at the 5,000 plus exoplanets to see evidence of civilizations in the atmospheres.((Gertner, 2022))

In the last ten years, astrophysicists have identified a few stars that could be evidence of a SETI structure around the star, such as a Dyson Sphere. For example, in 2015, using light

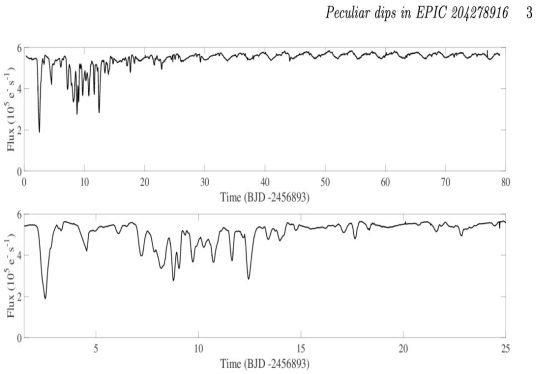
Figure 2: TABBY SETI Star – Credit: D. et al. ((2016))



curve data from Kepler, an astrophysicist identified a star that has a unusual transit light curve that could be evidence of a Dyson Sphere: KIC 8462852 is the Kepler star ID and the star is now named after the scientist who found it, Boyajian's Star. Above is the image of the Boyajian star's light curve ((D. et al., 2016)):

Similarly, in 2016, scientists identified another star using Kepler data that found an unusual light curve transit plots that could be evidence of SETI called EPIC 204278916.((Scaringi et al., 2016)) Like the light curve plot for the Boyajian star, the light curve transit plot for EPIC 204278916 is irregular ((Scaringi et al., 2016)):

Figure 3: EPIC 204278916 SETI Star – Credit: Scaringi et al. ((2016))



In 2019, using Kepler data, scientists identified another anomalous transit that could be evidence of SETI called EPIC 204376071, a star just 440 light-years from Earth, that unusually dimmed by up to 80 percent for an entire day. ((Rappaport1 et al., 2019))

By comparison to the light curve for the Boyajian star, EPIC 204278916, and EPIC 204376071, below at Figure 5 [2] there is a light curve plot of a typical transit from TESS data that is already on the 5,000 TOI list. As you can see, the transits in this light curve are symmetrical.

One of the biggest differences between the three anomaly light curves that could be SETI –

[2]All uncredited images in this report are originals I created.

Figure 4: EPIC 204376071 SETI Star – Credit: Rappaport1 et al. ((2019))

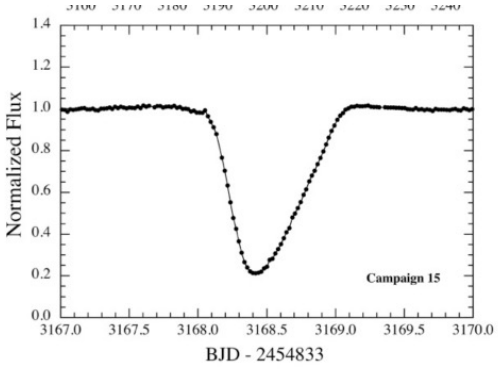
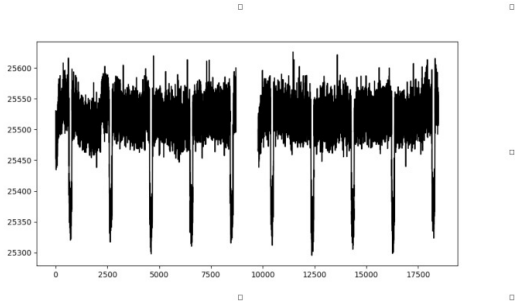


Figure 5: Typical TOI List Star



the Boyajian star, EPIC 204278916, and EPIC 204376071 – and the 5000 stars that are on the TOI list, is that almost all the light curve transits on the TOI list have transits where the light value generally never reduces less than five percent of the medium light value. I ran a Python program that I created to go through TOI lists and determine what the median is for the lowest 20 light values, lowest 100 light values, and lowest 500 light values. The exoplanets that are on the TOI list do not have transits that reduce the light more than 5 percent of the median value of the light from such stars.

c. AI, GANS, and Machine Learning.

In the last decade, as astrophysics equipment such as the TESS satellite have obtained massive data from space, there has been an influx of data that scientists have to analyze:

> In the last decades, the exponential growth of data has changed the way we do science. New and increasingly sophisticated astronomical facilities, both ground-based and spaceborne, produce massive amounts of data and they will be able to reach in few years a production rate in the order of petabytes per year. This data tsunami, both in terms of volume and velocity, will bring Astronomy in the big data era.((Garofalo et al., 2017))

One tool scientists have started using to analyze massive data has been computer program machine learning (ML):

> Machine learning was defined in 1959 by Arthur Samuel as "the field of study that gives computers the ability to learn without being explicitly programmed". ML allows to uncover hidden correlation patterns through an iterative learning by sample data (or past experiences) instead of being explicitly programmed. Common classes of problems that ML algorithms can solve are classification, regression, clustering, and outlier detection. These algorithms have been successfully used in astrophysics to solve different tasks. ((Garofalo et al., 2017))

In recent years, astrophysicists have used various types of machine learning programs to analyze massive TESS data and that of its predecessor, Kepler. See, e.g., ((Osborn et al., 2020)) ((Ofmana et al., 2021)) ((Friedman, 2020)) ((Tu1 et al., 2022)) ((Ansdell et al., 2018)) ((Shallue1 and Vanderburg, 2018)). While some of these astrophysicists have employed various types of "convoluted neural networks" (CNN) to characterize TESS and Kepler data, none of them have employed a new breed of CNN called GANs.

Since 2014, a type of machine learning program called GANs has appeared that is a powerful form of artificial intelligence to analyze large data.((Goodfellow et al., 2014)). Originally theorized in 2014 in a paper by the computer scientist Ian J. Goodfellow, the concept of GANS is as follows:

> In the proposed adversarial nets framework, the generative model is pitted against an adversary: a discriminative model that learns to determine whether a sample is from the model distribution or the data distribution. The generative model can be thought of as analogous to a team of counterfeiters, trying to produce fake currency and use it without detection, while the discriminative model is analogous to the police, trying to detect the counterfeit currency. Competition in this game drives both teams to improve their methods until the counterfeits are indistinguishable from the genuine articles.

Since Goodfellow's paper, computer scientists have created various GANS machine learning programs, including probably the most famous GANS program: creating photos of people that do not exist and which humans cannot figure out are not real:

> The ability of AI to generate fake visuals is not yet mainstream knowledge, but a new website — `www.ThisPersonDoesNotExist.com` — offers a quick and persuasive education.
>
> The site is the creation of Philip Wang, a software engineer at Uber, and uses research released last year by chip designer Nvidia to create an endless stream of fake portraits. The algorithm behind it is trained on a huge dataset of real images, then uses a type of neural network known as a generative adversarial network (or GAN) to fabricate new examples.((vincent))

In 2020, graduate students at the Massachusetts Institute of Technology developed a type of GANs Machine Learning program, called Orion-ML with a pipeline called TadGan that very efficiently goes through "time series" data to find anomalies:

> In this work, we introduce a novel GAN architecture, TadGAN, for the time series domain. We use TadGAN to reconstruct time series and assess errors in a contextual manner to identify anomalies. We explore different ways to compute anomaly scores based on the outputs from Generators and Critics. ((Geiger et al., 2020))

The significance of this Orion-ML/TadGan program, which is still in alpha phase of development, is that the model does not have to be trained on data that it already knows are anomalies in time series. The learning is unsupervised:

> Unlike with supervised time series anomaly detection, we do not have any previously identified "known anomalies" with which to train and optimize the model. Rather, we train the model to to learn the time series patterns, ask it to detect anomalies, and then check whether the detector identified anything relevant to end users.((Geiger et al., 2020))

3 Methodology And Data

For my research, my methodology included using a version of the Orion-ML program that used a TadGan pipeline. to AI Lab at MIT. This Orion-ML program in turn runs on TensorFlow, a machine learning program that works in Python Language that Google created:

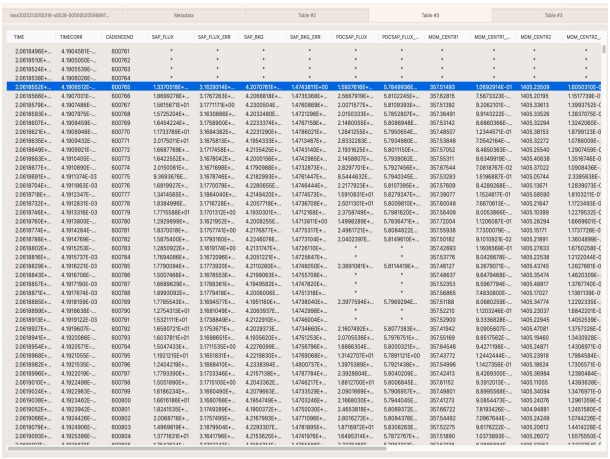
> TensorFlow is an end-to-end open source platform for machine learning. It has a comprehensive, flexible ecosystem of tools, libraries and community resources that lets researchers push the state-of-the-art in ML and developers easily build and deploy ML powered applications.((Google))

I setup my program in PyCharm, a program on my home Macintosh computer that allows one to use Python language to create programs. ((jetbrainsGoogle)) I installed the Orion-ML library and other libraries that work with Python. In particular, I needed a Python program that works with FITS files in which the TESS data I am using is stored. As the MAST website explains:

> The astronomical community has adopted the Flexible Image Transport System (i.e. FITS) format as the default standard for the exchange of data between institutions. The FITS file format is platform independent, supported by many institutions, and endorsed by both NASA and the IAU. For these reasons FITS is the recommended file format for archiving data at STScI. A description of the MAST data format recommendations can be found in the MAST Data Format Guidelines document. ((, FITS))

This is what a FITS file looks like, i.e. many columns and thousands of rows:

Figure 6: Typical Tess FITS File



To work with FITS files, I installed a Python Library called Astropy. Astropy. Astropy allowed my Python program to open a FITS file and access the data and then feed it to the TadGan machine learning program to train a model.

Here is what my Python code looks like, particularly a function called `"data_fix3"`, that cleans up the FITS file data so it can be fed into the TadGan machine learning part of the program:

```
#fix light values by normalizing and putting the medium value in nan rows
def data_fix3(name):
    tess_data = []
    new_array = []
    file_address = name
    open_file = fits.open(file_address, ignore_missing_end=True)
    cut_one = open_file[1].data
    time_stamp = 0
    for i in cut_one:
        time_stamp = time_stamp + 1
        row = [int(time_stamp), float(i[7]), float(i[5]), float(i[14]), float(i
    [16])]
        tess_data.append(row)
    cut_one_array = np.array(tess_data)
    median_value1 = np.nanmedian(cut_one_array[:,1], axis=0)
    median_value2 = np.nanmedian(cut_one_array[:,2], axis=0)
    median_value3 = np.nanmedian(cut_one_array[:,3], axis=0)
    median_value4 = np.nanmedian(cut_one_array[:,4], axis=0)
    array_df = pd.DataFrame(cut_one_array)
    array_df[2].fillna(value=median_value2, inplace=True)
    array_df[3].fillna(value=median_value3, inplace=True)
    array_df[4].fillna(value=median_value4, inplace=True)
```

```
    final_array = np.array(array_df)
    for i in final_array:
        if (pd.isnull(i[1])):
            row = i[0], i[1], (i[2] / median_value2), (i[3] / median_value3), (
    i[4] / median_value4) + 0.15
        else:
            row = i[0], i[1]/median_value1, (i[2]/median_value2), (i[3]/
    median_value3), (i[4]/median_value4) + 0.15
        new_array.append(row)
    columns = ['timestamp', 'value', 'back', 'cent1', 'cent2']
    cleaned_array = pd.DataFrame(data = new_array, columns=['timestamp', 'value
    ', 'back', 'cent1', 'cent2'])
    cleaned_array['timestamp'] = cleaned_array['timestamp'].apply(int)
    test_data = cleaned_array
    return test_data;
```

I downloaded from the MAST website the TOI list csv file which had approximately 5,000 stars with exoplanets. By creating a TadGan model that would train on the time series of files of the 5,000 exoplanet stars, my model would learn what normal exoplanet lightcurves could be. By learning what normal exoplanet transits light curves are, the TadGan Model could then find in the 50 Sectors (1 million FIT Files) such files that have transits that are normal or not normal compared to the transits on the TOI list that have exoplanets.

I set the TadGan Orion-ML program to train on three epochs for each of the files in the TOI list, which has about 5,000 TESS files. An epoch is one training run on the data in a TESS file. Three epochs then means three training runs through the data in a TESS file. My program started training the model on May 28, 2022 and by June 3, 2022 the training on the TOI list had been completed. I saved the TadGan model as a "pickle" file so I could use it again in Orion-ML to hunt for anomalies. It is important to note that this analysis would have been much faster if my TensorFlow program had used GPUs on the workstations I employed.[3] But the problem was that Orion-ML is still in alpha so it uses an older version of Python, 3.6, that did not work with the programs that worked with the GPUs on the workstation. So it took several days to train the model because TensorFlow just used the slower CPUs.

This is what the code in my program that calls the Orion-ML machine learning program TadGan looks like. Note, my program tells TadGan how many "epochs" to use when training on the TOI FITs files and other parameters to use for the training.

```
from orion import Orion
```

[3]I split the work between my mentor's workstation at MIT, which I accessed remotely with PyCharm, and an iMac computer at home.

```
hyperparameters = {
                        "mlprimitives.custom.timeseries_preprocessing.
    time_segments_aggregate#1": {
                            "interval": 1,
                        },

                        "mlprimitives.custom.timeseries_preprocessing.
    rolling_window_sequences#1": {
                            "window_size": 100,
                            "target_column": [0, 1, 2, 3]
                        },
                        "orion.primitives.tadgan.TadGAN#1": {
                            'epochs': 3,
                            'verbose': True,
                            'input_shape': [100, 4]
                        }
            }
orion_test = Orion(
    pipeline='tadgan',
    hyperparameters=hyperparameters
    )
```

My next step was to use the TadGan model, which I saved as a "pickle" file, to be loaded into a similar PyCharm program, go through the 50 Sectors of TESS files (approximately 1 million files) and provide an anomaly score for each light curve compared to the normal transit light curves that are on the TOI list. Significantly, I only wanted my trained TadGan Model to score light curves that are not "pulses". "Pulse" light curves are where the light value increases in value compared to the medium light value of the star. Because the model was trained on transit light curves, the program would treat all stars with pulses as anomalies. Because of that, I added code to the program that prevents files with pulses from being fed to the TadGan Model. To achieve that, my program eliminated light curves that had more than 20 light values that were greater than 3 percent of the medium light values in the time series. This is the code that counts how many light values are greater than 3 percent of the median:

```
cut_one_array = np.array(open_file)
second_col = cut_one_array[:, [1]]
first_col = cut_one_array[:, [0]]
first_col_nozeros = first_col[~np.all(first_col == 0, axis=1)]
count_values = np.count_nonzero(second_col > 1.03, axis= 0)
```

The Orion-ML program works by using TadGan to assign a "severity" score for the time

series. Time series that are assigned a high severity score are likely to include a large anomaly. When my program was set to find and plot time series having a severity score greater than a certain value (e.g. greater than 0.90), its TESS file would be added to a csv spreadsheet with the name of the TIC file. In addition, the program, using Astroquery, downloaded the following data to add to the spreadsheet from the MAST and SIMBAD database website:

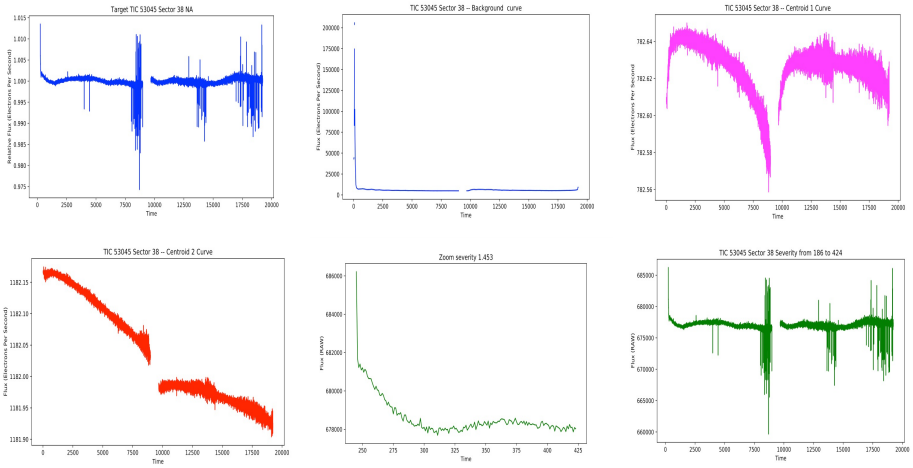
- TESS magnitude of the source (Tmag)
- V magnitude of the source (Vmag)
- TESS magnitude of the source (Tmag)
- V magnitude of the source (Vmag)
- The parallax values (Plx)
- The type of star (Lumclass), i.e. Dwarf, Giant, or Subgiant
- Radial Velocity Value `(Rv_Value)`

de Données astronomiques de Strasbourg ((CDS)) NASA ((b))[4] The significance of this additional data for each TESS file is that such data can help astrophysicists further analyze the star found by the program.

My program then added to the spreadsheet the start time and end time of the anomaly time series where the severity was larger than 0.90 and the actual amount of severity. My program also added to the spreadsheet the time that the TESS file was analyzed by the TadGan model. This program also plotted the light curves into individual pdfs with several plots. The first plot was the light curve over time, then the amount of background light over time plotted to see if it affected the light curve plot, then the two centroid values as separate plots over time to see if the light curve was affected, and then zoom plots of the areas on the time series where there was significant "severity," i.e. anomaly in the light curve. (The significance of the background light and centroid plots is that they can help the astrophysicist figure out if the the main light curve is a false positive that was created by light from the background or problems with the lens for the TESS camera. The zoom plot helps the astrophysicists look closer at the light plot where the main anomaly is located)

Here is a typical pdf of all the plots in a typically found file. For example:

[4]All of this data came from the MAST website except for the Radial Velocity data which came from the SIMBAD website.

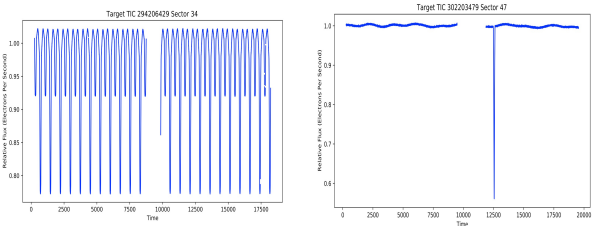


Skew Value

Also, my program (using the Python statistics library) ran a `"skew"` calculation on the light value column for each file. Skew is :

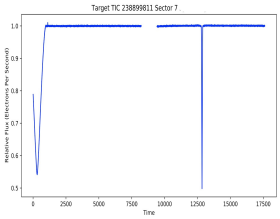
> For normally distributed data, the skewness should be about zero. For unimodal continuous distributions, a skewness value greater than zero means that there is more weight in the right tail of the distribution. The function skewtest can be used to determine if the skewness value is close enough to zero, statistically speaking.

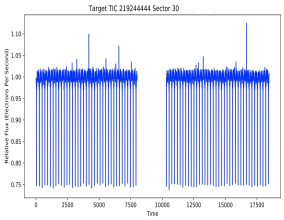
A skew value close to zero indicates symmetry in the light curve while a high number, either positive or negative, indicates asymmetry. For example, the light curve on the left below has a skew value that is closer to zero: -1.78. The light curve on the right has a value further away from zero indicating it is asymmetrical: -16.19



Difference

My program ran another calculation for each found file called "Difference" where the program summed up the total light values on the left side of the light curve and subtracted the total sum of the values from the right side of the light curve. If the symmetry of the light curve was heavily off, one side of the light curve sum should be greater than the other so any Difference value that was close to zero indicates symmetry and values further from zero indicated non-symmetry. For example, see below, the asymmetrical light curve on the left has a Difference value away from zero. The symmetrical light curve value on the right has a Difference value close to zero.





Ave Lowest

My program ran another calculation for each found file called `"ave_Lowest"` finding the lowest light values in the light curve. The program calculated three values: the medium value of the 20 lowest light values, the medium value of the 100 lowest light values, and the medium value of the 500 lowest light values. For each file, the program then added up those three values and found the average of those three and added that value to the spreadsheet. The value generated by this code gives one a sense of the lowest values in the light curve as the average light was plotted around the number 1. Here is the relevant code from my program:

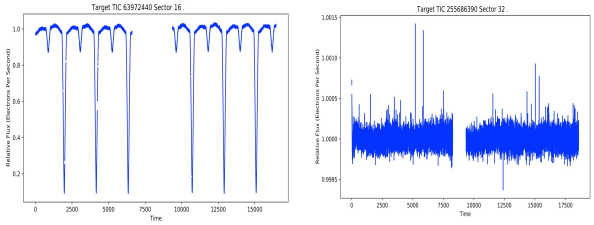
```
#finds the median of the lowest 20 light values
Lowest_20 = toi_dataframe.nsmallest(20, ["value"])
low_twenty = np.array(Lowest_20["value"])
median_low20 = np.nanmedian(low_twenty)

#finds the median of the lowest 100 light values
Lowest_100 = toi_dataframe.nsmallest(100, ["value"])
low_100 = np.array(Lowest_100["value"])
median_low100 = np.nanmedian(low_100)

#finds the median of the lowest 500 light values
Lowest_500 = toi_dataframe.nsmallest(500, ["value"])
low_500 = np.array(Lowest_500["value"])
median_low500 = np.nanmedian(low_500)

#creates an average of the three lowest values
ave_Lowest = (median_low20 + median_low100 + median_low500)/3
```

Below is an example of how `ave_Lowest` value indicates the depth of a transit of a light curve plot. On the left is a light curve plot with a `ave_Lowest` value close to zero: 0.12. On the right is a light curve plot with a `ave_Lowest` value close to 1 . The plot on the left has transits reduce in value much greater than the plot on the right that has transits that barely reduce in light value than the average light coming from that star.



Plotting Out Light Curves With Transits lower Than than 20 Percent of Median

The program also included Python code that looked into a found anomaly file and identified, plotted and put into a separate csv spreadsheet those files that had light curves with transits going further down than 20 percent of the median value. The Python code I created for this task was a simple "if" statement:

```
1 if ave_Lowest < 0.8:
```

Speed of My Program

Because my program could not use the GPUs on the workstations (because Orion-ML in alpha is using Python 3.6), the program worked much slower than when TensorFlow can use GPUs. The Tadgan model generally took approximately 1 minute to analyze a file. To make the analysis work faster, I created 14 copies of my program and had them work in parallel focusing on different sectors in the 50 sectors. To go through all one million files (and using only the CPUs), the 14 programs worked from approximately from July 7, 2022 to August 31, 2022.

4 Results

Anomaly Files Found by TadGan Machine Learning

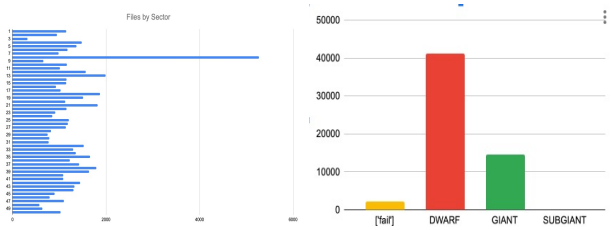
The spreadsheet that my program created listed 62,444 files from the approximately 1 million TESS files in the 50 Sectors that are transit anomalies compared to the transits in the TOI list that trained the TadGan model.

Each of the 62,444 files was given a "severity" score over 0.90. Some of these 62,444 files were multiple files from different sectors for the same TIC number, i.e. for the same star. Comparing these 62,444 files to the files already found on the TOI list, shows that 1923 of those files are already on the TOI list. [5]

The two charts below show information about the 62,444 anomaly files identified by the TadGan program. The chart on the left shows the number of anomalous files with "severity"

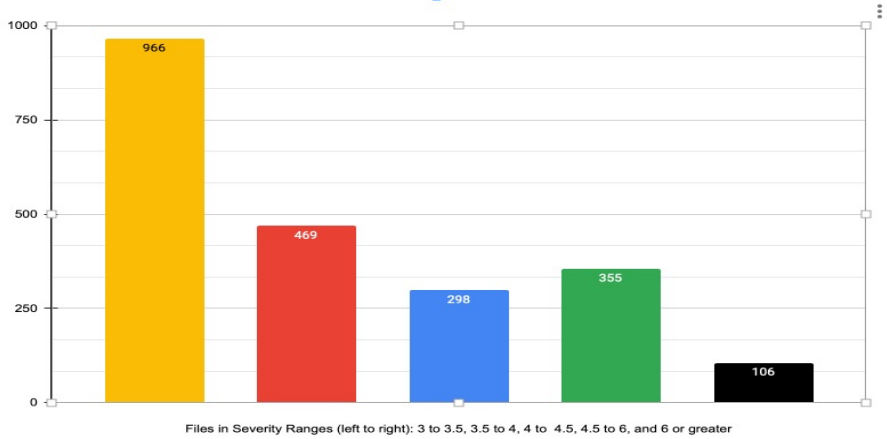
[5] So although the TadGan program and model was trained on the 5,000 TIC number related files that are on the TOI list, some of the found "anomaly" stars were already stars on the TOI list. This could mean that this TadGan model is not entirely accurate in finding anomalous files compared to the TOI list. But that could be because the files on the TOI list might be so different different from each other that the TadGan Model cannot train to believe each of those TOI files are normal. In training, the TadGan model may view some of the files on the TOI list to be anomalous compared to the majority of stars on that list.

over 0.90 that the TadGan model found in each of the 50 sectors. Interestingly, generally the number of files in each sector is less than 2,000, except for some reason Sector 8, which is anomalous, as it contains 5,269 anomalous files. Sector 3 contains the fewest anomalous files with only 319. Further analysis is needed to find out why the Tadgan Model found fewer anomalous files in Sector 3 compared to the massive amount in Sector 8. The chart on the right shows the distribution of "Lumclass" values. The vast majority of stars are "Dwarf" class, with approximately on third of that amount being "Giant" class stars. Only one file is a "Subgiant" class star. (Where "fail" is indicated means the program was not able to retrieve "lumclass" data for that file from MAST.)



High "Severity" Scores

Given that there are 62,444 TESS files with severity scores greater than 0.90, I decided to look closer to the files that have very large severity scores. I created a program that found and plotted out from the main spreadsheet into separate folders the files that had severity scores between 3 and 3.5, 3.5 and 4, 4 and 4.5, 4.5 and 6 and greater than 6. As the chart below shows, there were very few files (black bar) with "severity" scores greater than 6, only 106. In the lower "severity scores, between 3 and 3.5 (yellow bar), there were the most totaling 978.

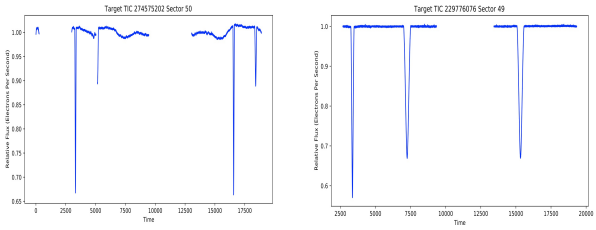


As discussed below, this methodolgy found some possible SETI candidates as well as other possible anamolous stars.

Lower Than 20 Percent From Median Light From the Star

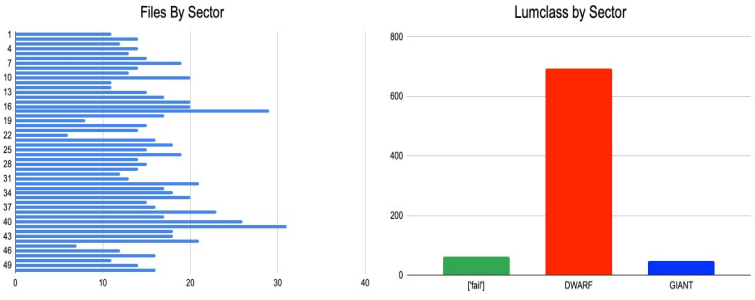
Another method my program used to find possible SETI candidates was to focus in the 62,444 anomaly files for light curve plots with transits that had light values reducing more than

20 percent of the median light value. The program found only 801 files from the 62,444 files that met that criteria. Here are two plots of two of the files from the 801 files as example. As you can see, they have transits where the light values reduce more than 20 percent. The one on the left has transits that reduces nearly 35 percent and the one on the right has a transit reducing more than 40 percent of the average light value for that star.



In both the master spreadsheet and the sub spreadsheet for the the 801 files that have light data reducing more than 20 percent of the median light value, I added a column that compared the files with the list of stars that are on the TOI list. The point of this column was to see if the files that were found were already on the list of TOI that are candidates of exoplanets.

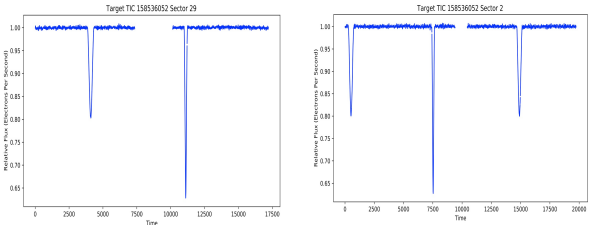
In the hunt for stars that are SETI candidates, my procedure looked at anomaly transit light curves that have 1) transits reducing more than 20 percent, and 2) being asymmetrical. Of the 62,444 anomalous files, only 801 files have light curve plots that have transits where the light reduces more than 20 percent from the median light of the star. The chart below on left shows the distribution of these 801 files in the various sectors that meet the first category – light values for transits that reduce more than 20 percent of the median light value. The most files of that category found in any sector are 31 in sector 41. The fewest are 6 in sector 21. Most sectors have between 10 and 20 files that fit that category. The chart on right shows the type stars in the 801 files found. The vast majority are "dwarf" stars while only a few are "giant" stars.



Significantly, none of the 801 files match up with TIC numbers that are already on the TOI list of found exoplanets.

In terms of the best SETI candidates, those would be the files with light plots with 1) transits lower than 20 percent, and 2) asymmetrical. Of the 801 files, there are 367 files that have skew values that are 6 or larger from than 0. But these light curves are not necessarily

asymetrical as the light curves could be over more than 1 sector. For example, this light curve with a skew for one file over 6, is in two sectors and when you show the two plots next to each other you can see together they are symetrical.

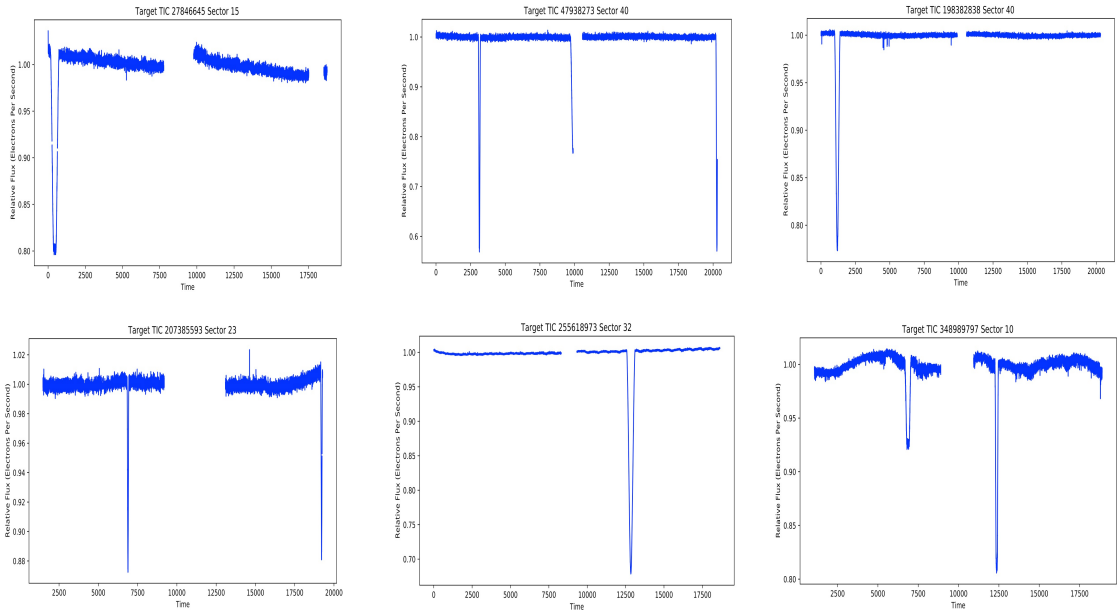


SETI Candidates

Interesting SETI candidates were found through the two methodologies, (1) identifying files with very high "severity" scores; and (2) identifying files with light values reducing more than 20 percent and high skew values.

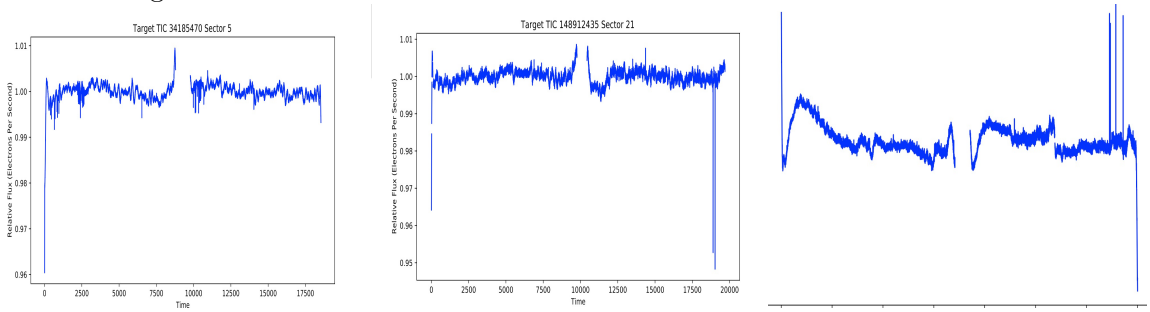
Under the first method, examining the 106 files with the larges severity scores (over 6), most of the light curve plots did not appear to be good SETI candidates as the transits did not generally reduce very much in value from the median value light, i.e. more than 5 percent. However, a few of the 106 plots with the highest severity did have transits that had more than 5 percent decline in light value from the median light value and unusual curves that could be looked at closer. These, with their lower light values and non-symmetrical light curves, could be possible candidates for follow up finding if they are SETI stars.

Below are some examples that are promising SETI candidates with transits with light reducing significantly more than 5 percent of the median.

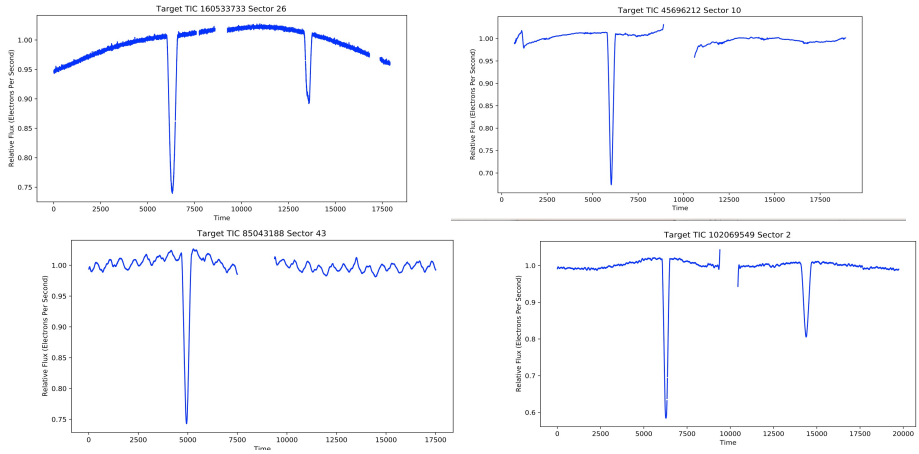


Even without the light value reducing more than 5 percent, there are light curves with high "severity" values that have asymmetrical light curves that could be possible SETI candidates,

like the following:

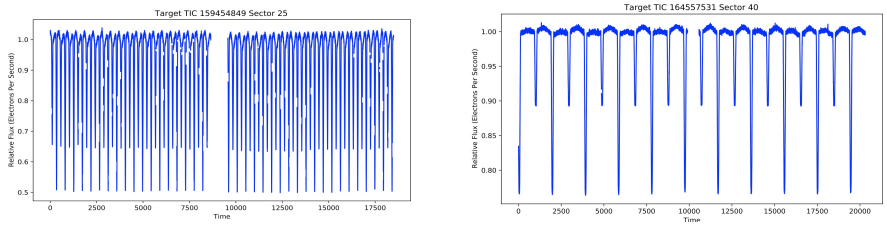


The second method, examining the 367 files with skew values greater than 6, also found TESS files that appear to be decent candidates for stars near SETI. The below light curve candidates have both (1) light values reducing more than 20 percent than median light value, and (2) are asymmetrical:.

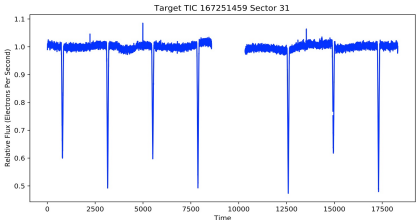


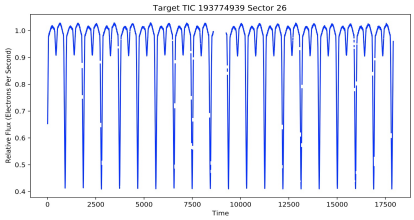
Massive Double Exoplanets or Binary Stars

These methodologies also found light curves that could be other interesting anomaly light curves. For example, these four light plots below are symmetrical but the repeated transits appear to be two different exoplanets as they reduce light value in different massive percentages and repeat. None of them are on the TOI list. They are very unusual compared to the TOI list as not only are they double, but the percentage of the star light they block is massive compared to the exoplanets on the TOI list. These light curves could be evidence of binary exoplanets, that are particularly large, and/or binary stars [6] that are rotating around each other and blocking light.



[6]The academic research indicates that binary stars systems can sometimes fool astronomers into thinking there are stars with exoplanets when the transits are being caused by binary stars.Bouma1 et al. ((2018))





5 Conclusion

This research shows that using a version of GANs unsupervised AI machine learning program with astrophysics data can help scientists find interesting anomaly data among massive data. The advantage of using Orion-ML and TadGan, the first time a GANs program was used on astrophysics data, is that it can create a model to score anomaly time series values in TESS data. This research shows it is possible to train a GANs model on TESS data, i.e. the 5,000 TOI files, and then use that GANs model to analyze 1 million TESS files. The Orion-ML pipeline my program used, TadGan, was able to cut down the 1 million Tess files to a fraction of 62,444 files that had "severity" scores of over 0.90. Using the Python program, these anomaly files could then be cut down even more based on either having (1) very large "severity" scores, such as over 4.5 or 6, and/or (2) low transit light values that are lower than 20 percent of the medium and not symmetrical. Like finding a needle in a haystack, my machine learning program was able to find at least twenty or more TESS files that could be decent candidates of SETI stars or other types of interesting candidates, such as large double exoplanets and binary stars. My program ultimately found approximately a dozen TESS files that astrophysicists can look closer at to see if they have SETI devices around such stars, like the Gabby star. In addition, my research found TESS files that could be evidence of stars with large double exoplanets that are not on the TOI list already, and possible star candidates that are binary stars that have not yet been identified. While the computer program took took several months to go through the 1 million TESS files, when Orion-ML is eventually updated to work with Python 3.8, this program will work much faster using the GPUs on a workstation. In addition, by tweaking the TadGan program, such as using more epochs, the GANs model can become more efficient in grading times series anomalies in TESS files. Finally, the program I created, using Orion-ML/TadGan, could be very help for going through and finding anomalies in other astrophysics data that have time series, such as the data from the new James Webb Telescope.

Acknowledgements

I would like to thank Dr. Tansu Daylan tremendously for his guidance, my research teachers, Dr. Macrae Maxfield and Ms. Stephanie Doire, and my supportive parents and sister.

References

M. Ansdell, Y. Ioannou, H. P. Osborn, M. Sasdelli, J. C. S. D. Caldwell, J. M. Jenkins, C. Räissi, and D. Angerhausen. Scientific domain knowledge improves exoplanet transit classification with deep learning. *The Astrophysical Journal Letters*, 869(1), 2018. URL `https://iopscience.iop.org/article/10.3847/2041-8213/aaf23b`.

Astropy. Astropy. URL `https://www.astropy.org`.

L. G. Bouma1, K. Masuda, and J. N. Winn. Biases in planet occurrence caused by unresolved binaries in transit surveys. *The Astronomical Journal*, 155(6), 2018. doi: https://iopscience.iop.org/article/10.3847/1538-3881/aabfb8.

T. S. B. D., D. LaCourse, S. A. Rappaport, D. Fabrycky, D. A. Fischer, D. Gandolfi, G. M. Kennedy, M. C. Liu, A. Moor, K. Olah, K. Vida, M. C. Wyatt, W. M. J. Best, F. Ciesla, B. C. k, T. J. Dupuy, G. Handler, K. Heng, H. Korhonen, J. Kova, T. Kozakis, L. Kriskovics, J. Schmitt, G. Szabo, R. Szabo, J. Wang, S. Goodman, A. Hoekstra, and K. J. Jek. Planet hunters x. kic 8462852 – where's the flux? *Monthly Notice of the Royal Astronomical Society*, 2016. doi: https://doi.org/10.48550/arXiv.1509.03622.

C. de Données astronomiques de Strasbourg (CDS). Simbad astronomical database - cds (strasbourg). URL `https://archive.stsci.edu`.

S. Developments. How to build a dyson sphere. URL `http://www.sentientdevelopments.com/2012/03/how-to-build-dyson-sphere-in-five.html`.

F. I. T. S. (FITS). Fits. URL `https://archive.stsci.edu/fits/`.

A. Friedman. The use of machine learning for exploring tess light curves. *NASA Center For Climate Simulation*, 2020. URL `https://www.nccs.nasa.gov/about-us/internships/intern-bios/adam-2020`.

M. Garofalo, A. Botta, and G. Ventre. Astrophysics and big data: Challenges, methods, and tools. *Proceedings of the International Astronomical Union, Cambridge University Press*, 12 (S325):pp.345–348, 2017. doi: https://doi.org/10.1017/S1743921316012813.

A. Geiger, D. Liu, S. Alnegheimish, A. Cuesta-Infante, and K. Veeramachaneni. Tadgan: Time series anomaly detection using generative adversarial networks. *IEEE International Conference on Big Data*, 2020. doi: https://doi.org/10.48550/arXiv.2009.07769.

J. Gertner. Technosignatures and the search for extraterrestrial intelligence, 2022. URL https://www.nytimes.com/2022/09/15/magazine/extraterrestrials-technosignatures.html.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. 2014. doi: https://doi.org/10.48550/arXiv.1406.2661.

Google. Tensorflow. URL https://opensource.google/projects/tensorflow.

jetbrainsGoogle. Pycharm program. URL https://www.jetbrains.com/pycharm/.

M. Kaufman. Technosignatures and the search for extraterrestrial intelligence. URL https://astrobiology.nasa.gov/news/technosignatures-and-the-search-for-extraterrestrial-intelligence/.

E. J. Korpela, S. M. Sallmen, and D. L. Greene. Modeling indications of technology in planetary transit light curves — dark-side illumination. *The Astrophysical Journal*, 809(139):13pp, 2015. doi: http://dx.doi.org/10.1088/0004-637X/809/2/139.

K. Masuda and K. Hotokezaka. Prospects of finding detached black hole–star binaries with tess. *THE ASTROPHYSICAL JOURNAL*, 883(2):pp.169–180, 2019. doi: https://doi.org/10.48550/arXiv.1808.10856.

NASA. Nasa's dart mission hits asteroid in first-ever planetary defense test, a. URL https://www.nasa.gov/press-release/nasa-s-dart-mission-hits-asteroid-in-first-ever-planetary-defense-test.

NASA. Barbara a. mikulski archive 8 space telescopes, b. URL https://archive.stsci.edu/astro/.

NASA. Kepler and k2, c. URL https://www.nasa.gov/mission_pages/kepler/launch/index.html.

NASA. Webb space telescope, d. URL https://webb.nasa.gov/index.html.

L. Ofmana, A. Averbuchc, A. Shliselbergd, I. Benaund, D. Segevd, and A. Rissman. Automated identification of transiting exoplanet candidates in nasa transiting exoplanets survey satellite (tess) data with machine learning methods. *New Astronomy*, 2021. URL https://arxiv.org/abs/2102.10326v2.

H. P. Osborn, M. Ansdell2, Y. Ioannou3, M. Sasdelli, D. Angerhausen, D. Caldwell, J. M. Jenkins, C. Räissi, and J. C. Smith. Rapid classification of tess planet candidates with convolutional neural networks. *Astronomy and Astrophysics*, 633(A53), 2020. doi: https://doi.org/10.1051/0004-6361/201935345.

S. Rappaport1, G. Zhou, A. Vanderburg, A. Mann, M. Kristiansen, K. Olah, T. Jacobs, E. Newton, M. Omohundro, D. LaCourse, H. Schwengeler, I. Terentev, D. Latham, A. Bieryla,

M. Soares-Furtado, L. Bouma, M. Ireland, and J. Irwin2. Deep long asymmetric occultation in epic 204376071. *Monthly Notice of the Royal Astronomical Society*, 2019. URL https://arxiv.org/pdf/1902.08152.pdf.

S. Scaringi, C. F. Manara, S. A. Barenfeld, P. J. Groot, A. Isella, M. A. Kenworthy, C. Knigge, T. J. Maccarone, L. Ricci, and M. Ansdell. The peculiar dipping events in the disk-bearing young-stellar object epic 204278916. *Monthly Notice of the Royal Astronomical Society*, 2016. URL https://arxiv.org/pdf/1608.07291.pdf.

C. J. Shallue1 and A. Vanderburg. Identifying exoplanets with deep learning: A five-planet resonant chain around kepler-80 and an eighth planet around kepler-90. *The Astrophysical Journal Letters*, 155(2), 2018. URL https://doi.org/10.3847/1538-3881/aa9e09.

D. to AI Lab at MIT. Orion. URL https://pypi.org/project/orion-ml/.

Z.-L. Tu1, Q. Wu1, W. Wang1, G. Q. Zhang1, Z.-K. Liu1, and F. Y. Wang. Convolutional neural networks for searching superflares from pixel-level data of the transiting exoplanet survey satellite. *The Astrophysical Journal*, 935(90), 2022. URL https://iopscience.iop.org/article/10.3847/1538-4357/ac7f2c.

J. vincent. This persondoesnotexist.com uses ai to generate endless fake faces. URL https://www.theverge.com/tldr/2019/2/15/18226005/ai-generated-fake-people-portraits-thispersondoesnotexist-stylegan.