

Stage I: Idea Proposal

Research Question:

Does the star have a planetary system that we can find through planets transit light curve?

While space missions and telescope send us millions of gigabytes of raw data needs to be analyzed, astronomers spend years to analyze and study a few portion of the data, a machine learning model would be extremely useful to examine the large chunk of untouched data such as exoplanet data. This will aid future mission designers and scientists.

Datasets:

1- NASA Transit data observed by TESS (<https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=TOI>)

2- All ever observed transit EDT (<http://var2.astro.cz/ETD/>)

3-Kepler labled time series data (<https://www.kaggle.com/keplersmachines/kepler-labelled-time-series-data>)

Machine Learning approach and method:

Supervised machine learning/ Neural networks

Sedra Abulbasir

Stage II: Paper Selection

Group nickname:

Planetary Hunters

Group members:

1. Asma Abdulsamad Abdulaziz
2. Hajar Hameed Raslan
3. Sedra Abdulghaffar Abdulbasir

Selected research question:

Does the star have a planetary system that we can find through the planet's transit light curve?

Selected papers:

<i>Title</i>	<i>Author</i>	<i>Publication Venue</i>	<i>URL</i>
--------------	---------------	------------------------------	------------

<i>A MACHINE LEARNING TECHNIQUE TO IDENTIFY TRANSIT SHAPED SIGNALS</i>	Susan E. Thompson ^{5,1,2} , Fergal Mullally ^{2,3} , Jeff Coughlin ² , Jessie L. Christiansen ⁴ , Christopher E. Henze ³ , Michael R. Haas ³ , and Christopher J. Burke ^{2,3}	The Astrophysical Journal	https://iopscience.iop.org/article/10.1088/0004-637X/812/1/46/meta
<i>AUTOMATIC CLASSIFICATION OF KEPLER PLANETARY TRANSIT CANDIDATES</i>	Sean D. McCauliff ¹ , Jon M. Jenkins ² , Joseph Catanzarite ² , Christopher J. Burke ³ , Jeffrey L. Coughlin ² , Joseph D. Twicken ² , Peter Tenenbaum ³ , Shawn Seader ² , Jie Li ³ , and Miles Cote ²	The Astrophysical Journal	https://iopscience.iop.org/article/10.1088/0004-637X/806/1/6/meta
<i>Searching for Possible Exoplanet Transits from BRITE Data through a Machine Learning Technique</i>	Li-Chin Yeh ¹ and Ing-Guey Jiang ²	Publications of the Astronomical Society of the Pacific	https://iopscience.iop.org/article/10.1088/1538-3873/abbb24/meta
<i>A Search for Multiplanet Systems with TESS Using a Bayesian N-body</i>	Kyle A. Pearson	The Astronomical Journal	https://iopscience.iop.org/article/10.3847/1538-3881/ab4e1c/meta

<i>Retrieval and Machine Learning</i>			
<i>Exoplanet Detection using Machine Learning</i>	Abhishek Malik, Benjamin P. Moster, Christian Obermeier	Earth and Planetary Astrophysics	https://arxiv.org/abs/2011.14135
<i>A convolutional neural network (CNN) based ensemble model for exoplanet detection</i>	Ishaani Priyadarshini & Vikram Puri	Earth Science Informatics	https://link.springer.com/article/10.1007/s12145-021-00579-5

Stage III: Literature Review

Given your selected papers (Step 2), skim the papers again. For each chosen paper, answer the following questions:

Asma Abdulsamad Abdulaziz (3rd, 4th paper)
Hajar Hameed Raslan (1st, 2nd paper)
Sedra Abdulghaffar Abdulbasir (4th, 5th paper)

First paper: A MACHINE LEARNING TECHNIQUE TO IDENTIFY TRANSIT SHAPED SIGNALS

1) What venue was the paper published in? Who are the authors? How many times has the paper been cited?

Published in The Astrophysical Journal ,the authors are Susan E. Thompson^{5,1,2}, Fergal Mullally^{2,3}, Jeff Coughlin², Jessie L. Christiansen⁴, Christopher E. Henze³, Michael R. Haas³, and Christopher J. Burke^{2,3} . The paper been cited 54 times.

2) What problem are the authors trying to solve?

Discuss and implement a machine-learning technique to determine whether a signal looks like a transit. develop more sophisticated metrics to cull-out those detections that do not look like transits.

3) Why is the problem important ?

A machine learning model helps Scientists to remove more than 90% of the non-transiting signals while retaining more than 99% of the known planet candidates.

4) What mathematical notation do the authors use and what does this notation mean? This is often most useful represented as a table.

<i>notation</i>	<i>notation mean</i>
x_i	<i>data vector</i>
A^T	$n \times N$ transformation matrix.
W_{ij} ,	<i>Data vectors i and j</i>
y_i	<i>dimensions</i>
$y_i = A^T x_i$,	
$\sum_y (y_i - y_j)^2 W_{ij}$. (1)	

5) What datasets did the authors use (if any)?

The Kepler data used to search for Earth-sized planets around distant stars

6) What machine learning models did the authors use (if any)?

k-nearest-neighbors

7) What evaluation metrics did they use? What do these metrics mean?

Using Lpp metric ,A linear programming problem consists of a linear function to be maximized or minimized subject to certain constraints in the form of linear equations or inequalities.

8) What baselines did they compare against?

Local Linear Embedding

9) How is your project similar and different from this paper?

Similar in the problem domain (astronomy, exoplanet detection). They used the same data set (Kepler labelled time series dataset). We will possibly use a different machine learning method other than the stated in the paper.

Second paper: Automatic Classification of Kepler Planetary Transit Candidates

1) What venue was the paper published in? Who are the authors? How many times has the paper been cited?

Published in The Astrophysical Journal, the authors are Sean D. McCauliff¹, Jon M.

Jenkins², Joseph Catanzarite², Christopher J. Burke³, Jeffrey L. Coughlin², Joseph D. Twicken², Peter Tenenbaum³, Shawn Seader², Jie Li³, and Miles Cote². The paper has been cited 69 times.

2)What problem are the authors trying to solve?

This paper presents the application of machine-learning techniques to the classification of the exoplanet transit-like signals present in the Kepler light curve data. Transit-like detections are transformed into a uniform set of real-numbered attributes.

3)Why is the problem important ?

A machine learning model would be extremely useful to examine the large chunk of untouched data such as exoplanet data. This will aid future mission designers and scientists.

4)What mathematical notation do the authors use and what does this notation mean?

This is often most useful represented as a table.

<i>Notation</i>	<i>notation mean</i>
-----------------	----------------------

X	<i>attributes of the TCEs.</i>
$ X $	<i>number of attributes</i>
$X = \{x_1, \dots, x_A\}$	<i>the set of a TCE's attributes</i>
I	<i>refers to a split made on the ith attribute</i>
Y	<i>refers to the set of all classes to which a TCE might belong</i>
$y_i \in Y$	<i>A particular label</i>
\mathcal{L}	<i>training set.</i>
$N = \mathcal{L} $	<i>The size of the training set</i>
$\mathcal{L}_l = \mathcal{L}_{x_i < s}$	<i>one subtraining set contains all of the TCEs with an attribute less than the split</i>
$\mathcal{L}_r = \mathcal{L}_{x_i \geq s}$	<i>all the TCEs greater than or equal to the split</i>
H	<i>classification function</i>

5)What datasets did the authors use (if any)?

Kepler mission datasets represented as Light Curves.

6)What machine learning models did the authors use (if any)?

Randomforest ,Decision Trees.

7)What evaluation metrics did they use? What do these metrics mean?

$$h(\text{MES}_{\min}/\text{MES}_{\max}, \Delta_{\text{centroid}}) = \begin{cases} \text{classify TCE as NTP, if } \frac{\text{MES}_{\max}}{\text{MES}_{\min}} < s_1 \\ \text{classify as PC, if } \Delta_{\text{centroid}} < s_2 \\ \text{classify as AFP, otherwise} \end{cases} \quad (1)$$

Multiple Event Statistic (MES), Δ_{centroid} mean out-of-transit centroid in arcseconds.

8) What baselines did they compare against?

k-nearest neighbors (k-NN) and naive Bayes.

9) How is your project similar and different from this paper?

Similar in the same problem and used the same data set We will possibly use a different machine learning method.

Third paper: A Search for Multiplanet Systems with TESS Using a Bayesian N-body Retrieval and Machine Learning:

1) What venue was the paper published in? Who are the authors? How many times has the paper been cited?

Publication Venue: The Astronomical Journal
 Authors: Kyle A. Pearson
 7 times of citations.

2) What problem are the authors trying to solve?

The search for multi-planet systems.

3) Why is the problem important?

Because of the gravitational influence of companions, transiting exoplanets in multiplanet systems have non-Keplerian orbits, which can cause transit times and durations to vary.

4) What mathematical notation do the authors use and what does this notation mean? This is often most useful represented as a table.

simultaneously maximizing the likelihood of a transit model and nonlinear correction	$F_{\text{obs}} = (a_0 + a_1 * t + a_2 * t^2) F_{\text{transit}}.$
The next mid-eclipse or mid-transit time	$t_{\text{next}} = n_{\text{orbit}} * P + T_{\text{mid}},$
A nonlinear ephemeris is used to predict the transit times of multiple planets in a system	$t_{1,\text{next}} = f_{\text{Nbody}}(n_{\text{orbit}}, M_*, M_1, P_1, e_1, \Omega_1, P_2, M_2, e_2, \Omega_2) + T_{1,\text{mid}},$

5) *What datasets did the authors use?*
Not included in the paper.

6) *What machine learning models did the authors use?*
Neural network and Bayesian algorithms.

7) *What evaluation metrics did they use? What do these metrics mean?*
The evaluation matrices were only the accuracies of the algorithms. The accuracy of a machine learning model is a metric for determining which model is the best at identifying relationships and patterns between variables in a dataset based on the input, or training, data.

8) *What baselines did they compare against?*
The comparison was done by using multiple parameters and plotting the accuracy of the algorithm.

9) *How is your project similar and different from this paper?*
In terms of similarity, the similarity is in searching in the planet system. In terms of difference, the difference is in the research question and applying a different machine learning algorithm.

Fourth paper: Searching for Possible Exoplanet Transits from BRITE Data through a Machine Learning Technique:

1) *What venue was the paper published in? Who are the authors? How many times has the paper been cited?*
Publication Venue: Publications of the Astronomical Society of the Pacific
Authors: Li-Chin Yeh¹ and Ing-Guey Jiang
24 times of citations.

2) *What problem are the authors trying to solve?*
The search for the exoplanet transits which are possible.

3) *Why is the problem important?*
Because after the Kepler Space Telescope discovered thousands of exoplanet candidates, the transit method became one of the major players among many detection techniques.

4) *What mathematical notation do the authors use and what does this notation mean? This is often most useful represented as a table.*
Not included in the paper.

5) *What datasets did the authors use?*
BRITE Data.

6) *What machine learning models did the authors use?*
Convolutional Neural Network.

7) *What evaluation metrics did they use? What do these metrics mean?*
The evaluation matrices were accuracy, reliability, and completeness. The errors are the standard deviations of the versions, and the accuracy, reliability, and completeness values are the corresponding median values.

8) *What baselines did they compare against?*
They implemented multiple models which are Model P12, Model P23, Model P34 and Model P45 and compared the accuracy, reliability, and completeness between them.

9) *How is your project similar and different from this paper?*
In terms of similarity, the similarity is in searching in the planet system. In terms of difference, the difference is in the research question and applying a different machine learning algorithm.

Fifth paper: Exoplanet Detection using Machine Learning

1) *What venue was the paper published in? Who are the authors? How many times has the paper been cited?*

The paper was published at arXiv preprint archive. The article's authors are: Malik, Abhishek, Benjamin P. Moster, and Christian Obermeier. This paper has been cited once (It has not been published in a journal yet, it is a preprint).

2) *What problem are the authors trying to solve?*

Detecting the extrasolar planet transit signals by classifying the data coming from space missions such as Kepler and TESS.

3) *Why is the problem important ?*

First, the datasets coming from space are huge, we receive millions of lightcurves every week. In this specific problem domain, we can find possible candidates of earth-like habitable planets if we analyze this big data observed of the universe around us.

4) *What mathematical notation do the authors use and what does this notation mean? This is often most useful represented as a table.*

Unfortunately, the authors of this work did not provide any mathematical notations. However, I find it good if I include some of the interesting results illustrated in the paper's figures.

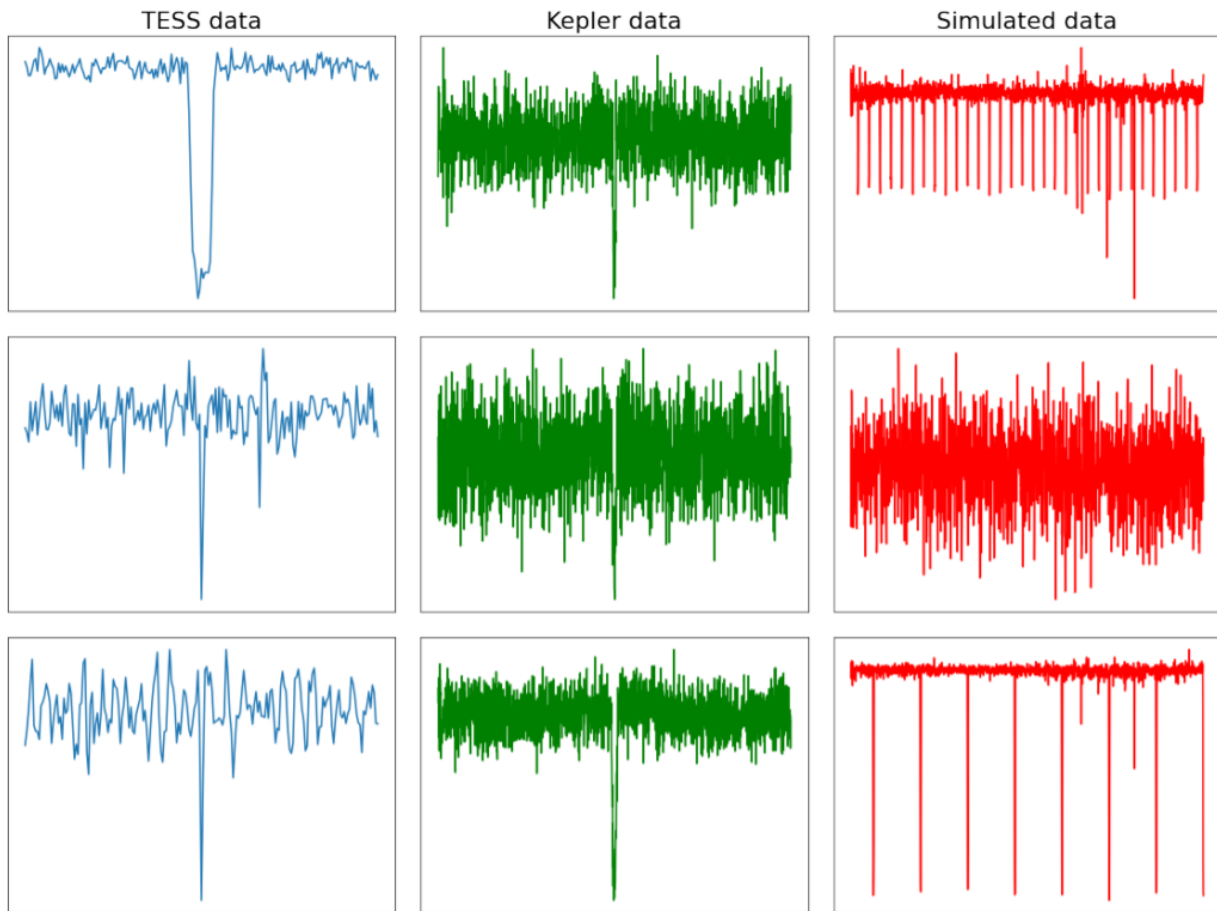


Figure 1. A comparison figure between the noisy lightcurve (data) of Kepler and TESS and the cleaned simulated lightcurve of the transit signals.

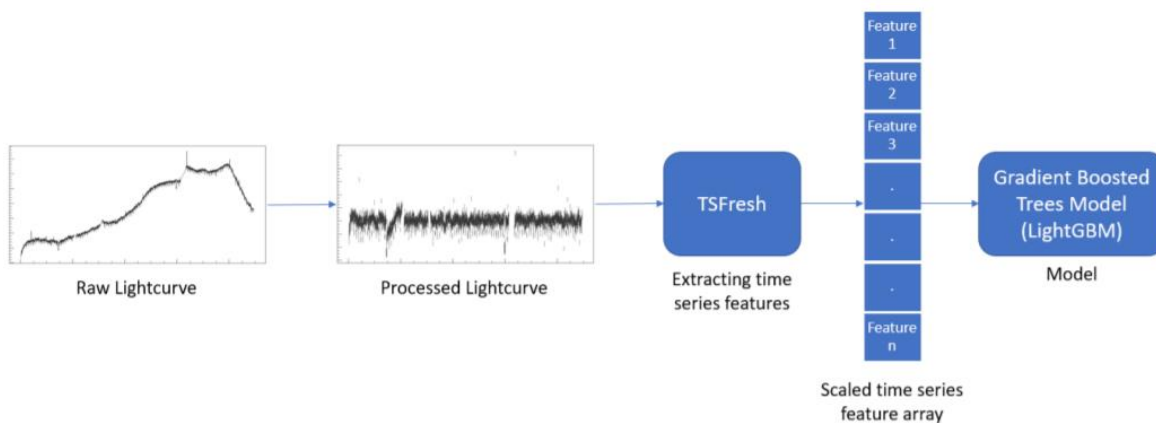


Figure 2. *The workflow of the machine learning model used in detecting transiting signals of exoplanets that could be useful for our project.*

5) What datasets did the authors use (if any)?

The researchers used NASA's Kepler mission datasets, along with Kepler's successor, Transiting Exoplanet Survey Satellite (TESS) dataset.

6) What machine learning models did the authors use (if any)?

Binary Classification, Highly Efficient Gradient Boosting Decision Tree (as a tool).

7) What evaluation metrics did they use? What do these metrics mean?

According to the paper, these are the evaluation metrics:

1- AUC: The area under the curve: demonstrates the percentage of times a planet candidate is ranked higher than a false positive.

2- Recall: shows the percentage of of total planet candidates in the validation set is recovered.

3- Precision: Out of the total predicted planet candidates by model, what % are actually "planet candidates".

4- Accuracy: the percentage of the correctly classified samples, including both planet candidates and false positives.

These four parameters can help identify better candidates of potential exoplanets or planetary systems.

8) What baselines did they compare against?

In the first stage, they used K2 photometry and additional injected transits from the simulated dataset as the baseline, they processed the lightcurve, extracted the features in time series, before feeding Kepler and TESS data to the LightBGM model .

9) How is your project similar and different from this paper?

The datasets we were planning to use are the same as the researchers used to train their model. They used a relatively simple machine learning model rather than building a complex and expensive deep learning model, which is similar to our plan. These datasets contain high amounts of noise, processing the lightcurve and cleaning the data and the removal of the noise and outliers they did, could be useful for our project since we are using the same datasets. We will possibly use a different machine learning method, using additional datasets

Sixth paper: A convolutional neural network (CNN) based ensemble model for exoplanet detection

1) What venue was the paper published in? Who are the authors? How many times has the paper been cited?

The paper was published in Earth Science Informatics (2021) journal. The researchers are : Priyadarshini, Ishaani, and Vikram Puri. Their work has been cited six times.

2) What problem are the authors trying to solve?

The active astronomy research topic of Detecting extrasolar planets, finding possible earth-like habitable planets using the transit method.

3) Why is the problem important ?

Analyzing the big unused astronomy data, helping physicists, cosmologists, scientists, researchers, academicians, industry experts, and machine learning experts, since the detection of exoplanets has been and still burning problem for decades.

4) What mathematical notation do the authors use and what does this notation mean? This is often most useful represented as a table.

<i>notation</i>	<i>notation mean</i>
<i>TN</i>	<i>True Negative</i>
<i>TP</i>	<i>True Positive</i>
<i>FN</i>	<i>False Negative</i>
<i>FP</i>	<i>False Positive</i>
$\frac{TP + TN}{TP + TN + FP + FN}$	<i>Accuracy</i>
$\frac{TP}{TP + FP}$	<i>Precision</i>
$\frac{TP}{TP + FN}$	<i>Sensitivity</i>

$\frac{TN}{TN + FP}$	Specificity
----------------------	-------------

5) What datasets did the authors use (if any)?

The paper did not mention all data sets used, since it is more like a review paper, but they focused on Kepler Labelled Time Series Data.

6) What machine learning models did the authors use (if any)?

Decision Trees, Support Vector Machines, Logistic Regression, Random Forest Classifier, Multilayer Perceptron (MLP), Convolutional Neural Network(CNN), Ensemble-CNN Model.

7) What evaluation metrics did they use? What do these metrics mean?

According to the paper, these are the parameters:

- 1- Accuracy: number of correctly predicted data points with respect to the total number of data points.(correctly classified exoplanets)
- 2- Precision: number of relevant observations with respect to retrieved observations.
- 3- Sensitivity (recall): ratio of actual positive events that got predicted as positive.
- 4- Specificity: ratio of actual negatives that got predicted as negative.

8) What baselines did they compare against?

They set Decision Trees, Support Vector Machines, Logistic Regression, Random Forest Classifier, Multilayer Perceptron (MLP), Convolutional Neural Networks as the baseline algorithms. Ensemble-CNN model to compare it with different machine learning models which they found has the best performance.

9) How is your project similar and different from this paper? (4-5 sentences)

The paper did a helpful comparative analysis of several machine learning models that we studied in this subject. Therefore, we might choose one of the best performing models stated in this paper. Also, most of the reviewed models have been fed with the same dataset we will use. We will try different datasets alongside with the Kepler dataset. We will try to explore more machine learning models other than the ones stated in the paper.

Stage IV: Fortnightly Reports

After the literature review, you will submit a report. The purpose of the report is to make sure you are on track with your project.

The reports should answer the following questions:

1) What activities did you do so far?

We learned the python programming language, then we tried to implement the basic libraries that the instructor provided. After that, we chose the kepler dataset among other datasets we provided in the idea proposal stage, then we attempted to train our labeled exoplanet dataset using a logistic regression model, then we tried to implement a Random Forest model using google colab. Finally, we trained at a Convolutional Neural Network(CNN) with our dataset. However, we could perform well due to the amount of issues we encountered.

We had daily meetings with the team members to perform the project tasks.

2) What preliminary results do you have?

The first logistic regression model training resulted in 77.5% accuracy which is relatively good for a first try.

Training the Random Forest Classifier resulted in 100% accuracy, which is obviously overfitting.

3) What issues have you encountered?

- 1- Understanding machine learning libraries (mostly pandas).
- 2- Programming with python is relatively new for us, we had issues in converting our thoughts through this language in the beginning.
- 3- Issues with printing the evaluation metrics and the confusion matrix function.
- 4- Issues in loading the datasets when using pandas.
- 5- Overfitting in the Random Forest model training.
- 6- Major issues while applying the Convolutional Neural Network (CNN) model.

4) What questions do you have?

- 1- How to increase the accuracy in the logistic regression model.
- 2- How to reshape the data to get better results.
- 3- How each attribute would change the results.
- 4- How to remove the overfitting in the Random Forest model.
- 5- How to graph and visualise the data to get a better understanding of the process.

5) What are concrete steps (for the next step)?

- 1- Using 2 models which are logistic regression and Random Forest and compare between their results and modify them to be realistic results.
- 2- Using other datasets to reach various results and benefit from them.

6) What are longer-term to do items?

- 1- Deploy the model in larger data instances to tune the model.

Test large amounts of data for the exoplanets datasets, and contribute in searching earth like exoplanet candidates.

2- Study more recent research papers to perform better data preprocessing and decrease the noise.

3- Refine the CNN model as the main model for our project.

4- Improve the project and take further steps to publish a scientific research paper, which would add value to the astronomy field and contribute to developing tools and models in this field, leading to solving major issues and obstacles that scientists and astronauts may encounter.

Stage V: Final Report

Abstract

The main mission of the Kepler space telescope was collecting data that would contribute to the discovery of stars with planets orbiting them. This data can be converted into a lightcurve, which is a graph of the observed star's light intensity over time. The changes in the light intensity sometimes occur in patterns that give useful information about the star. One pattern is the transit signal which can be an evidence of the existence of a planet. These are large amounts of data that might be challenging for astronomers to analyze. On the other hand, Artificial Intelligence can help accelerate the data analysis process to detect extraterrestrial planets (Exoplanets). In this work, we built different machine learning (ML) algorithms that can detect exoplanets from the Kepler time series dataset. We trained Random Forest, Logistic Regression models, and Convolutional Neural Network (CNN) models on the dataset, these models have been evaluated using accuracy, sensitivity, precision metrics. The Random Forest, Logistic Regression and CNN respectively scored 99.08%, 99.08%, 99.12% accuracy, 99.08%, 99.08%, 100% precision, 100%, 100% recall for all of them. Hence, they might be the best performing ML models that can be useful for astronomy and astrophysics researchers and other experts working in the extra terrestrial planet searching field.

Introduction

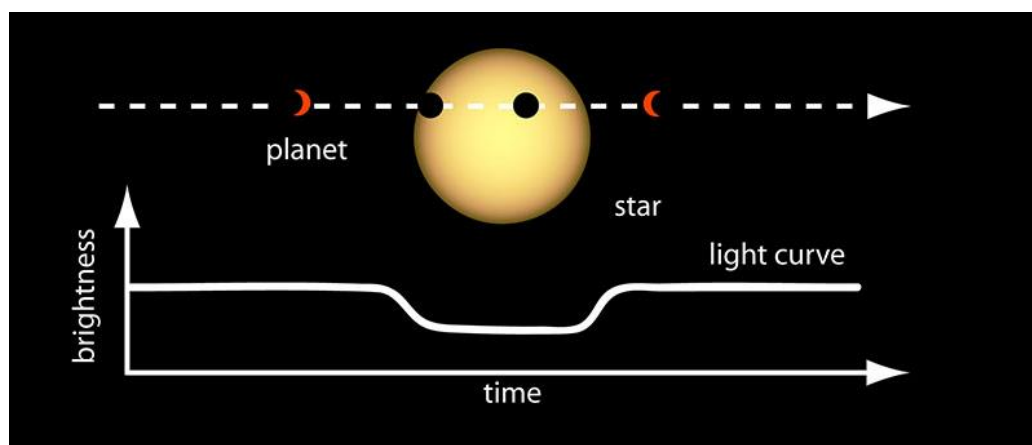
Recently, discovering the exoplanets in the near galactic stars has been an active research topic in astronomy and machine learning fields, aiming to find earth-like habitable planets that could be the humans' next station within the next decades or centuries. For that reason, in 2009, NASA sent the Kepler space telescope to scan the sky looking for planetary systems other than ours. By 2015, Kepler had discovered evidence of other potentially habitable planets within the Milky Way galaxy. It was a rocky Earth-like planet that has been named 'Kepler-452b'. It is known

to be the closest planet to Earth so far. There are many exoplanets discovered after that and have similar physical properties to the Earth. An enormous quantity of data is being transmitted from space missions and telescopes to earth, where astrophysicists and data scientists must analyze and investigate them. This process is complicated and time-consuming. Therefore there was a demand for automation of this process using data analysis tools and artificial intelligence solutions. (Pat Brennan, 2015).

A planetary system consists of one or more planets around a host star, which itself orbits in another orbit. Furthermore, planetary systems can be detected using a variety of techniques, including radial velocity, direct imaging, polarimetry, astrometry, transient photometry, gravitational microlensing, and transit time variation and other methods. Transient photometry is the most often employed technique for planet detection. It occurs when a planet passes in front of a star, resulting in a modest periodic drop in the star's light intensity, and is referred to as the transit signal. Once a transit signal is detected, we may use Kepler's third law of planetary motion to determine the planet's orbital size and the period required for the planet to orbit the star, in addition to the star's mass. As a result, the characteristic temperature of a planet may be determined using its orbital radius and the temperature of its star. With this in mind, the issue of whether the planet is habitable or not may be answered.

There are two types of transit; the first one is FACE-ON which considered the best transit and the second one is EDGE-ON which used with -radial velocity “The transit led to the monitoring of 530,000 stars in the Cygnus constellation and it has confirmed the existence of more than 2,600 exoplanets in 2018.”

Light curves are graphs that plot the brightness of an object over time Figure(1). They are an extremely useful instrument for studying objects whose brightness fluctuates over time, such as novae, supernovae, and variable stars. LightKurve is an easy-to-use tool for analyzing and visualizing time series data regarding the brightness of planets, stars, and galaxies. Originally developed to support astrophysicists using NASA's Kepler and TESS observatories, this package may also be used by amateur astronomers to examine light curves.



Figure(1). The transit of a planet causes a noticeable and gradual decrease in the flux of a star for a period of time (NASA Ames, 2021).

Related works

(Thompson, 2015) performed a k-nearest-neighbors machine learning model that yielded to 99% accuracy using Kepler mission data and (Kyle A. Pearson, 2019) trained a random forest machine learning model for the first time to classify the kepler mission dataset scoring 98.94 accuracy and 5.85% overall error rate. Both of the researches helped us to identify the testing methods that were used in this new research in a developed phase.

(Sean D. McCauliff, 2015) Sectors 1-3 of Kepler data set has been studied to search for transit timing variations (TTVs). Several artificial intelligence techniques, such as Convolutional Neural Network (CNN), with multi-input and multi-output systems have been conducted to quickly detect new planets by vetting non-transit signals before characterizing light-curve time series. In order to achieve other effective outputs and find other planets, an enhanced machine learning method is going to apply and study new sets of data.

(Li-Chin Yeh¹, 2020) The research aims to explore exoplanets moving around the lighted stars through photometric light curves that are analyzed from the BRITE satellites group by using Machine Learning algorithms and methods. We constructed several Convolutional Neural Network (CNN) models to find transit candidates based on different transit periods of time, which were trained by using synthetic transit signals combined with BRITE light curves until they were above 99.7% percentage of accuracy. The researcher used this method because it is very efficient and effective regarding the small number of selected transits that are possible to appear in the result. For instance the Machine Learning system had chosen only 2 systems "HD37465 and HD186882", which reached higher priority during the future observations processes. The result is valuable regarding focusing on 2 systems that would. (Abhishek Malik, 2020) Suggested using astrophysics approaches to improve the performance of the machine learning algorithms and tested them on a simulated dataset, yielding better results when compared to box least squares fitting. Training the Binary Classification models and Decision Tree on Kepler dataset resulted in 94.8% accuracy and 96% recall.

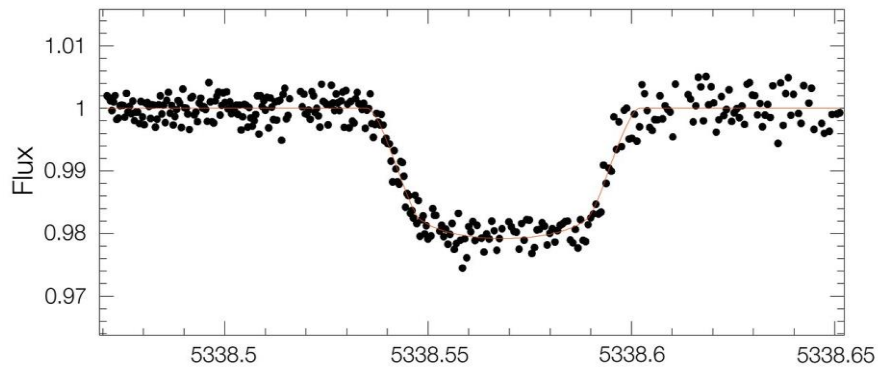
(Priyadarshini, 2021) Compared the performance of Decision Trees, Support Vector Machines, Logistic Regression, Random Forest Classifier, Multilayer Perceptron (MLP), Convolutional Neural Network (CNN) and Ensemble-CNN Model that have been trained on Kepler time series and K2 datasets, concluded that the Ensemble-CNN is the best performing model in exoplanet detection scoring 99.64% accuracy.

These research papers helped us a lot in our topic. The machine learning approaches that were used for detecting exoplanets aided us to select best performing models, more improved approaches to preprocess the dataset. AI made the processes of predicting and analyzing way easier than before by reducing time, cost, and efforts. Regarding these studies we realized that this is one

of most important topics in the Astronomy field and will add a huge value to this science by discovering new things in the field that contribute to the development of science and facilitate life for living organisms. Our main purpose is to explore new things about this wonderful planet.

Dataset

The Kepler mission sent us various information about the observed objects (stars) to search for Earth-sized planets around distant stars and designed to survey our region of the Milky Way galaxy. Each star has been observed over a period of time, mostly 8 hours. For this period of time, the luminosity or the flux of the star is the amount of energy emitted every unit of time. It changes due to several reasons, one reason can be the transit of a planet causing a decrease in the flux in a drop shaped pattern. See Figure (2).



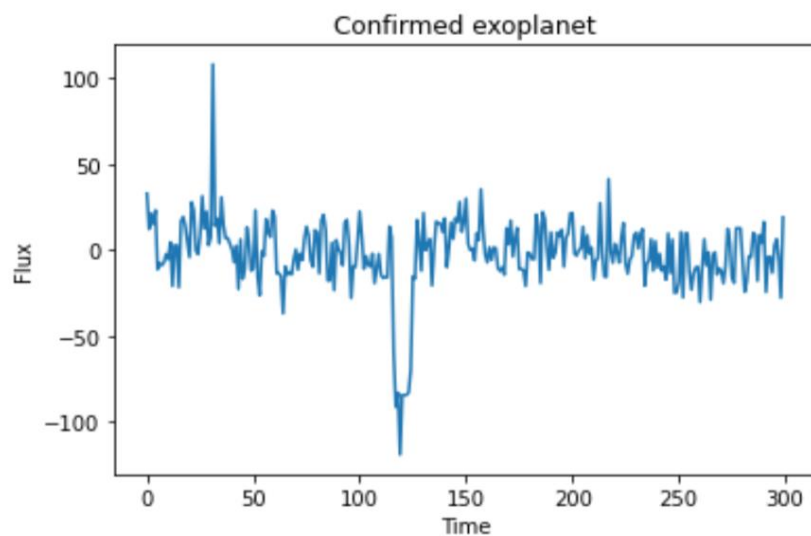
Figure(2).The transit of a planet causes a gradual drop in the light intensity of the host star within a period of time (ESO, 2021).

Our dataset has been retrieved from kepler mission archives, labelled and organized into a time series table, see Table(1). Column number 0 represents the predicted value or the label (1 , 2) not confirmed, confirmed planet respectively. While the remaining columns (1- 3197) represent the light intensity or the flux of the star observed in a period of time. Each row contains the brightness and the label of a certain object of interest (star) collected by Kepler.

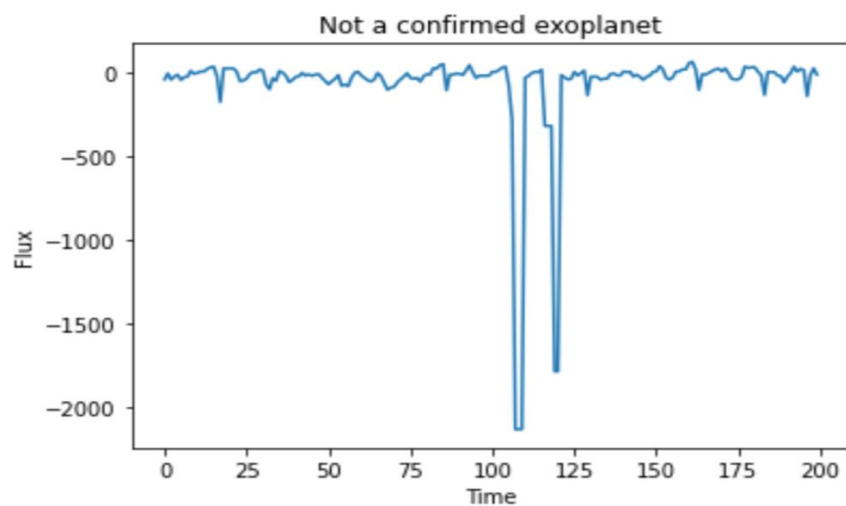
Table(1). First 5 examples of Kepler time series in the training dataset where the y is the label, x are the flux from 1 to 3197.

LABEL	FLUX.1	FLUX.2	FLUX.3	FLUX.4	...	FLUX.3193	FLUX.3194	FLUX.3195	FLUX.3196	FLUX.3197
2	93.85	83.81	20.1	-26.98	...	92.54	39.32	61.42	5.08	-39.54
2	-38.88	-33.83	-58.54	-40.09	...	0.76	-11.7	6.46	16	19.93
2	532.64	535.92	513.73	496.92	...	5.06	-11.8	-28.91	-70.02	-96.67
2	326.52	347.39	302.35	298.13	...	-12.67	-8.77	-17.31	-17.35	13.98
2	-1107.21	-1112.59	-1118.95	-1095.1	...	-438.54	-399.71	-384.65	-411.79	-510.54

The selected features for the training are 1-3197 columns (the light intensity). The label or the prediction value is the 0 column. For each instance in this time series dataset, we can create a light curve graph (figure(3), figure(4)).



Figure(3). a time series plotting of an object illustrates the transit signal as a sign of a confirmed planet before the data preprocessing (containing noise and outliers).



Figure(4). A time series plotting of an object illustrates a different type of a flux change that is not similar to a transit signal, hence is not a confirmed planet (containing noise and outliers).

Methods

The Kepler time series dataset has been preprocessed using methods similar to astrophysics data analysis. Then Random Forest, Logistic Regression and CNN models training has been performed on the dataset. Evaluation using accuracy, recall and precision metrics has been done. These models have been selected based on the previous work done on Kepler data. Then comparison to find the best performing algorithm to detect transit shaped signals has been performed.

Light Curve data preprocessing

We split data to training set and test set, but as seen in table (1), the data needed to be filtered and preprocessed in multiple ways, as it contains false positives, instrumental noise and outliers that makes it difficult to detect a transit-like signal.

We dropped the null rows and columns, and reshaped value, then performed scaling and transforming, then selected 3197 features,

The LightKurve library helps the astrophysicist to analyze Kepler and TESS time series data, by removing outliers, scattering, flattening, and folding the data. However, this tool uses a retrieved dataset from NASA archives. For this reason, we tried to create functions similar to the LightKurve library to preprocess our own dataset, these functions include: Fourier Transformation, Reducing Upper Outliers, Smoothing Filters, Normalization, Standardization.

We applied these functions by creating a class then creating an object containing the dataset before feeding the dataset to our machine learning models.

Training the models

I. Logistic Regression model:

Logistic regression is one of the most common machine learning algorithms used for binary classification. It predicts the probability of occurrence of a binary outcome using a logit function. It is a special case of linear regression as it predicts the probabilities of outcome using log function. In this case, the activation function (sigmoid) has been used to convert the outcome into categorical values.

Logistic Regression model training:

After the preprocessing stage, we built and trained several Logistic Regression models that iterated 6000 times, using Sklearn and Pandas libraries. Then we performed regularisation using GridSearchCV with $\text{penalty}=12$.

In the first training the scoring accuracy was 75.5%. After that we edited some of the preprocessing methods, accuracy increased to 77%. Finally, after using the light curve preprocessing class methods that imitates the astrophysics approaches the model scored 99% accuracy.

II. Random Forest Model:

Random Forest is a robust machine learning algorithm that can be used for a variety of tasks including regression and classification. It is an ensemble method, meaning that a random forest model is made up of a large number of small decision trees, called estimators, which each produce their own predictions. The random forest model combines the predictions of the estimators to produce a more accurate prediction.

Random Forest model training:

The stages of building this model:

- 1- Choosing random samples from the dataset.
- 2- Building multiple decision trees for all samples then get a prediction result from each decision tree.
- 3- Taking the result from each decision tree.
- 4- Choosing a predicted result depending on the most repetition result as the final prediction.

We implemented several random forest models using sklearn, pandas, numpy libraries, the result was scoring 100% accuracy which is an overfit, then we used the lightcurve data preprocessing class to remove any noise and outliers, the resulting accuracy went down to 99.8%.

III.Convolutional Neural Network Model:

The architecture of this type of artificial neural network follows a similar pattern to multilayer perceptrons. This type of NN is specifically designed to process pixel files, and since the data coming from Kepler are originally a series of pixel files converted into a lightcurve, it can be used in exoplanet detection.

Training the convolutional neural network model:

Before feeding the preprocessed dataset to the model, we had to scale and reshape it. Using Keras, TensorFlow, Sklearn and many other libraries, we build a convolutional neural network model. These libraries helped us flattening and max pooling the input then used ReLU and Sigmoid as the activation function, with 11 sized kernel. Editing some of the Lightcurve preprocessing class methods helped improve the accuracy from 99.08% to 99.12%.

Results and evaluation

The confusion matrix is helpful in evaluating the performance of a predictive model on unseen. it can be calculated in several steps:

1. Use part of the data (test dataset) as an expected outcome value (y_training_set).
2. Make a prediction for each row test dataset.
3. Compare the predicted result and true results.

These numbers are organized in a matrix. Rows refer to predicted class, columns refer to actual class.

Evaluation Metrics

After training the dataset, we used the testing set to evaluate the performance of our model. We used Accuracy, Precision, and Recall. The resulting metrics were significantly well performing.

Accuracy:

Accuracy defines how many times the machine learning model correctly classified the data, we calculate it using the formula:

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Where TP means True Positive, TN True Negative, FP False Positive, FN False Negative.

The CNN model scored **99.12%** accuracy.

The Logistic Regression model scored **99.08%** accuracy.

The Random Forest model scored **99.08%** accuracy.

Precision:

Precision defines the ratio of the true values and the predicted values, we calculate it using the formula:

$$\frac{TP}{TP + FP}$$

Where TP means True Positive, TN True Negative, FP False Positive.

The CNN model received **100%** precision.

The Logistic Regression model received **99.08%** precision.

The Random Forest model received **99.08%** precision

Recall:

Recall or Sensitivity is the ratio of the actual positive values and the predicted values, we calculate it using the formula:

$$\frac{TP}{TP + FN}$$

Where TP means True Positive, TN True Negative, FN False Negative.

The CNN model scored **100%** sensitivity.

The Logistic Regression model scored **100%** sensitivity.

The Random Forest model scored **100%** sensitivity.

Conclusion

As the search for Earth-like possibly habitable planets outside our solar system continues to grow, newer space missions to discover the sky emerge, creating huge amounts of raw data, increasing the demand to automate the data analyzation process. Using Artificial Intelligence could be the possible solution for this problem. After reviewing several papers that used machine learning methods to detect exoplanets, we obtained an insight of best performing models to analyze the Kepler mission dataset. In this work, we used astrophysics approaches to preprocess the data, then built Logistic Regression, Random Forest and CNN models, trained them on Kepler time series dataset. As a result, the performances of each were significantly upstanding. Logistic regression model scored 99% accuracy, Random Forest scored 99% accuracy and CNN reached 99.1% accuracy. Those promising results can indicate that continued testing and prediction can lead to discovering new unexpected extra terrestrial planets that could possibly be our next destination.

References

1. Thompson, S. E., Mullally, F., Coughlin, J., Christiansen, J. L., Henze, C. E., Haas, M. R., & Burke, C. J. (2015). A machine learning technique to identify transit shaped signals. *The Astrophysical Journal*, 812(1), 46. Retrieved from <https://iopscience.iop.org/article/10.1088/0004-637X/812/1/46/meta>
2. Yeh, L. C., & Jiang, G. (2020). Searching for Possible Exoplanet Transits from BRITE Data through a Machine Learning Technique. *Publications of the Astronomical Society of the Pacific*, 133(1019), 014401, Retrieved from <https://iopscience.iop.org/article/10.1088/1538-3873/abbb24>
3. Pearson, K. A. (2019). A Search for Multiplanet Systems with TESS Using a Bayesian N-body Retrieval and Machine Learning. *The Astronomical Journal*, 158(6), 243. Retrieved

from <https://iopscience.iop.org/article/10.3847/1538-3881/ab4e1c/meta>

4. Malik, A., Moster, B. P., & Obermeier, C. (2020). Exoplanet Detection using Machine Learning. arXiv preprint arXiv:2011.14135. Retrieved from <https://arxiv.org/abs/2011.14135>
5. Priyadarshini, I., & Puri, V. (2021). A convolutional neural network (CNN) based ensemble model for exoplanet detection. Earth Science Informatics, 14(2), 735-747. Retrieved from <https://link.springer.com/article/10.1007/s12145-021-00579-5>
6. Kepler. (2018). NASA. Retrieved from https://www.nasa.gov/mission_pages/kepler/overview/index.html
7. NASA, A Machine-Learning Algorithm Just Found 301 Additional Planets in Kepler Data (November 26, 2021). Universe Today <https://www.universetoday.com/153441/a-machine-learning-algorithm-just-found-301-additional-planets-in-kepler-data/>
8. Transit Method. (2021). Institute for Research on Exoplanets. <http://www.exoplanetes.umontreal.ca/transit-method/?lang=en>
9. LightKurve documentation. Retrieved from <https://docs.lightkurve.org/>
10. N.A.S.A.A. (2021c, March 4). Light Curve of a Planet Transiting Its Star. Retrieved from <https://exoplanets.nasa.gov/resources/280/light-curve-of-a-planet-transiting-its-star/>
11. P.B. (2015a). Finding Another Earth. Retrieved from <https://www.nasa.gov/jpl/finding-another-earth>