Using Machine Learning fo Exoplanet Classification

Juliana Wang

Abstract

With aims of discovering potential candidates and using that new found information to analyze the composition of the world as we know now, many efforts have been made to conduct research efficiently and accurately. With multiple methods for exoplanet detection such as shadow searching, the data produced from these methods still require interpretation to reach a conclusion (e.g. is there a dip in the light curve?), which is why machine learning has recently come into the scene of astronomy with its potential to be trained for image classification tasks, requiring only a couple of seconds to a couple of minutes to complete their task. This paper used convolutional neural networks (CNN), a type of machine learning specifically designed to classify images, and achieved an area-under-curve coverage of 0.91.

Keywords

exoplanet discovery, neural networks, computational astrophysics, machine learning

1 Introduction

In hopes of finding new habitable zones, new forms of life, and to better understand the origins of the universe, many scientists prioritize exoplanet research (Brennan, 2019), with scientists from NASA having confirmed the first near-Earth-size planet orbiting within the habitable zone of a star similar to ours (NASA, 2015). Since the 1990s NASA has successfully detected thousands of exoplanets by alternating in between different detection methods — radial velocity (difference in lightwave patterns as a result from a star's transit, 1075 planets discovered with this method), shadow searching (dim as a result of an exoplanet passing in front of its host star, 4151 discovered), direct imaging (observing light scattered from the exoplanet itself), gravitational microlensing (bending of light caused by the gravity of a passing body, 210 planets discovered) and astrometry (change in a star's position as a result from its planets orbit, 3 planets discovered) (NASA, n.d.). However, in order to read the data and to reach conclusions more efficiently, machine learning algorithms have recently been used to classify together images and visual patterns from observatories to identify the behavior of planetary motion (as the ones listed above).

As a way to organize data and conduct more thorough analysis to obtain results, data classification has been implemented to achieve this efficacy. With the invention of machine learning (ML) taking place, this set of algorithms have allowed models to learn through data sets without needing direct instruction, improving its results through training for a specific amount of time. This specific application of artificial intelligence (AI) has now been continuously developed and used throughout studies up until this day, to classify extraterrestrial and interstellar data in papers such as (Sturrock et Al., 2019, Jin et Al., 2022, Singh & Kumbhare, 2022) . By combining both ML and data classification to sort raw data sets (specifically related data sets derived from exoplanet observation) behaviors obtained from the observations make patterns and characteristics more distinct, allowing one to draw conclusions in a more efficient and orderly manner while training models and observing its operations and processes.

In this paper, a neural network (type of ML) is used to classify the Kepler data set [1] (a dataset which obtains 10000 exoplanet candidates (NASA & Bilogur, 2017)) to determine if the model deems fit to classify future candidates. Previously, since neural networks have been used and performed the best in exoplanet classification tasks (highest accuracy of 99.79% (Jin et Al., 2022)); this paper will furthermore carry out trials with differing layer amounts to determine which configuration provides the top performance.

2 Method

2.1 Neural Networks

A subset of artificial intelligence, neural networks are computer networks inspired by the structure of the human brain, with each node (denoted as a circle in Figure 1) representing a neuron, allowing it to process information through connections and passing data through layers/filters (Park & Lek, 2016). There are certain types of artificial neural networks: convolution neural networks (CNN), and recurrent neural networks (RNN). This paper will utilize the MLPClassifier, an artificial neural network. Different variations of layers will be modified within the model (e.g. 2-dimensional and three dimensional, differing layer numbers) to obtain the most accurate model. The accuracy will then be represented through a ROC curve (Receiver Operating Characteristic curve).

The structure of an artificial neural network further allows this specific algorithm to start learning with no prior information, and instead gives it an adaptive structure the more time it is trained. Aside from the input and output layers, neural networks contain hidden layers, and especially in the multi-layer perceptron classifier (MLP Classifier), a feedforward neural network (a type of artificial neural network) comprises an input layer, hidden layers and an output layer, a more basic neural structure. It does, however, use sigmoid neurons to process non-linear data efficiently (IBM, n.d.).

2.3 Data Definitions

In the architecture of our neural network, hyperparameters (the layer and dimension) of our neural network are modified by trial to identify the optimum number of layers for best accuracy, and the model type that has the highest area coverage under the ROC curve, and satisfactory

[1] https://www.kaggle.com/datasets/nasa/kepler-exoplanet-search-results

recall and precision rates (≥ 0.6) are analyzed. The Kepler dataset contains 50 columns and 9564 rows. However, only 'CONFIRMED' and 'FALSE POSITIVE' data points are used to test out the model, meaning that rows with 'CANDIDATE' status are eliminated, which fall under the koi_disposition row. (koi_disposition — the literature of an exoplanet, can be 'CANDIDATE', 'FALSE POSITIVE', 'NOT DISPOSITIONED' or 'CONFIRMED' (NASA & Bilogur, 2017)).

NAME	DEFINITION [2]
'koi_disposition'	values in the dataset that are under CANDIDATE, FALSE POSITIVE, NOT DISPOSITIONED or CONFIRMED
'koi_period'	Time in between planetary transits
	Visible distance between the star's surface and its exoplanet
'koi_duration'	The duration of the observed planet's transit
'koi_depth'	Dip in stellar light/Dim in stellar lightness. Typically computed from data
'koi_prad'	Radius of the observed planet
'koi_teq'	Rough temperature of the planet
'koi_insol'	Stellar equilibrium temperature
'koi_model_snr'	Depth of stellar flux caused during transits
'koi_steff'	Photospheric temperature of the star
'koi_slogg'	Acceleration from stellar gravity (in base-10)
'koi_srad'	Photospheric radius of the star

(**Table 1.** Definitions of terms used in the Kepler Dataset)

[2] https://exoplanetarchive.ipac.caltech.edu/docs/API_kepcandidate_columns.html

Furthermore, empty rows are deleted and only a certain amount of data points are fed to the network, as shown below. (The data points fed to the Neural Network were numbers under columns 'koi_disposition', 'koi_period', 'koi_impact', 'koi_duration', 'koi_depth', 'koi_prad', 'koi_teq', 'koi_insol','koi_model_snr','koi_steff', 'koi_slogg','koi_srad').

These columns were classified as “Important” in the algorithm —(columns that did not seem to be relevant to the discovery process were dropped) and represent transits and patterns obtained from the observatory. As mentioned initially, these patterns are a representation of planetary behavior (dip/change in light), and the numbers will allow the machine to understand what values fall under the categorization of “CONFIRMED” and “FALSE POSITIVE”. By tracking these two classes, it will help depict the accuracy of the model due to its ability to differentiate the estimations with the results of the labeled data set. Furthermore, due to the presence of some empty rows, instead of filling them up they will be removed with the function df_important.dropna, leaving around 4000 rows of data, which will be sufficient (considerably big amount, 4599 of 9564 rows (48%)).

- In the data collection, the following metrics were used to evaluate the performance of the model:

Precision (measures the accuracy of predictions):

$$\frac{True\ Positives}{True\ Positives + False\ Positives} \qquad (1)$$

And to calculate the recall (accounts for the total amount of predictions):

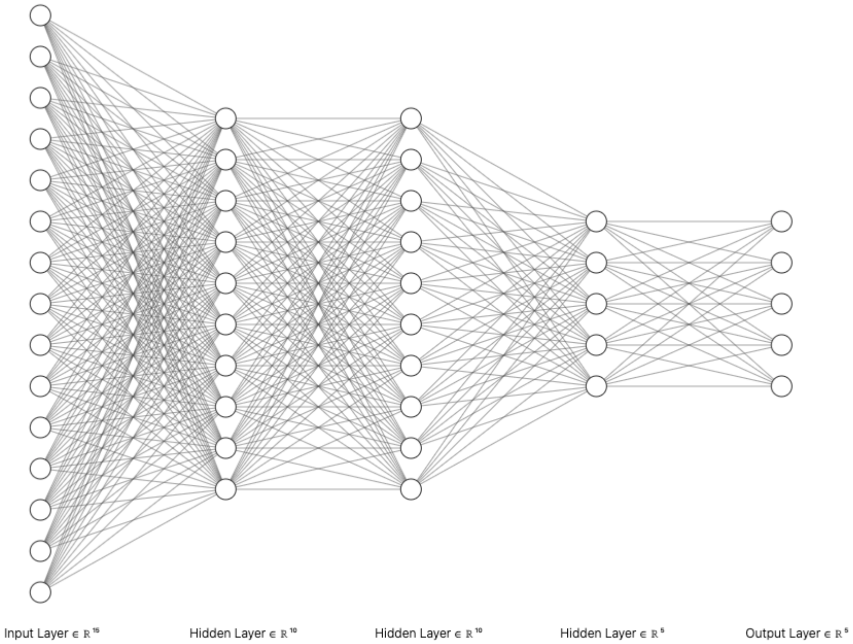
$$\frac{True\ Positives}{True\ Positives + False\ Negatives} \qquad (2)$$

ROC Curve — A receiver operating characteristic (abbreviation ROC, also known as ROC curve) is a visual plot that shows the performance of a binary classifier system (such as a neural network machine learning algorithm) by showing its accuracy through the area covered under the curve line formed (typically a high recall rate produces better results). It is calculated by inputting the true positive rate against the false positive rate.

The layers will be modified throughout trials (starting by (50,50)), and follows the visual structure in figure 1.

Index	koi_period	koi_impact	koi_duration	koi_depth	koi_prad	koi_teq	koi_insol	koi_model_snr	koi_steff	koi_slogg	koi_srad
25%	2.24	0.22	2.53	184.60	1.52	559.00	23.35	14.70	5320.00	4.21	0.83
50%	8.51	0.58	3.91	507.45	2.68	928.00	176.51	30.75	5779.50	4.44	1.00
75%	36.18	0.92	6.43	2775.10	25.45	1496.25	1201.17	123.10	6126.00	4.54	1.37
count	7316.00	7016.00	7316.0	7016.00	7016.00	7016.00	7057.00	7016.00	7016.00	7016.00	7016.00
max	1071.23	100.81	138.54	1541400.00	200346.00	14667.00	10947554.55	9054.70	15896.00	5.28	229.91
mean	58.82	0.79	5.87	30620.12	129.97	1148.59	8485.67	326.63	5727.71	4.30	1.78
min	0.24	0.0	0.11	0.80	0.08	92.00	0.02	0.00	2661.00	0.05	0.12
std	121.08	3.67	6.97	92873.98	3519.62	898.33	160221.87	891.67	825.22	0.44	6.20

(**Table 2.** Summary statistics for all final set of used features)



(**Figure 1.** Neural Network model that displays the best accuracy during trials with 5 dimensions and layers (300,200,200,100,100) (divided by 20) Generated through https://alexlenail.me/NN-SVG/index.html)

3 Results

3.1 Process

An artificial neural network (MLP Classifier) is built by using the scikit learn python library [3]. The activation function used in the nodes of the model is the ReLU function, and the solver used was the adam solver (Kingma, 2014). It trains on two arrays (n_samples and n_features) and another y array (n_samples). Both contribute to the training set of the algorithm. After the training the model will be fit for testing, and once the results are obtained a library imported from sklearn (3.3 Model Evaluation [4]) is used to code out a graphical plot for the ROC curve.

After the output is displayed, the results from the end of the trial are plotted on a table, which displays that only Trial 15 reached the desired statistics. Each trial is run 3 times.

3.2 Data Collection Table

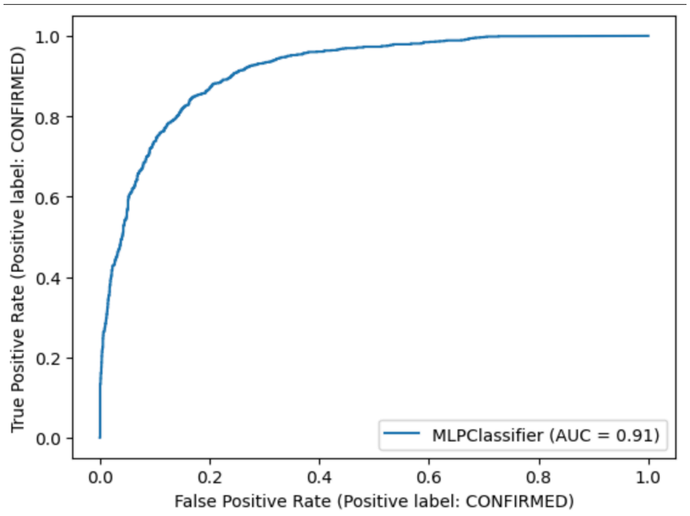
[3] https://scikit-learn.org/stable/modules/neural_networks_supervised.html
[4] https://scikit-learn.org/stable/modules/model_evaluation.html#roc-metrics

16 trials were conducted to figure out the most accurate layer amount.

The trial starts with two dimensions. The first three trials had layer numbers that added up to 100, and to avoid overfitting instead of adding a larger layer amount another dimension would be included. Trials 4-7 are then increased by increments of 100, along with trials 8-11 (layer amounts would start from 100 and gradually increase until a fallout). The rest of the layers only have the layer amount adjusted in a limited number of dimensions to avoid overfitting caused by excessive layer and dimension numbers.

3.3 ROC Curve Figure and Data Interpretation

Trial 15 was the only round that demonstrated the highest ROC curve area coverage percentage (Shown in figure 2) with precision and recall rates above 0.6, while Trials 14, 12, 11, 9, 6 and 5 obtain a high area coverage in the ROC graph, however recall and precision rates for both classes did not consistently meet 0.6 or above. Despite Trial 11-15 retaining a similar AUC percentage, on Trial 16 the area coverage starts to fall off, potentially meaning that the machine learning algorithm might have picked up noise from the input data. Therefore, in this paper CNN with configurations of 3 dimensions (with high layer amounts such as (400,400)), 4 dimensions (with low layer amounts such as (100,100,100,100)) and 5 dimensions (with low layer amounts such as (300, 200, 200, 200, 100)). Anything slightly higher might cause overfitting (the picking up of noise) and anything lower did not seem to perform as high as trials 11-15 in terms of area coverage in ROC Curve. For class “CONFIRMED”, the model seemed to miss 40%, while for class “FALSE POSITIVE” the model missed 5%, misclassified 15% of “CONFIRMED” and 17% of “FALSE POSITIVE”.



(**Figure 2** - Roc Curve Output of Trial 15)

Trial #	Amount of Layers	Recall - FALSE POSITIVE	Precision - FALSE POSITIVE	Recall - CONFIRMED	Precision - CONFIRMED	ROC Curve area coverage
1	(50,50)	0.65	0.97	0.95	0.57	0.87
2	(70,30)	0.95	0.80	0.50	0.84	0.89
3	(40,60)	0.72	0.95	0.92	0.61	0.88
4	(200,200)	0.68	0.94	0.91	0.58	0.88
5	(300,300)	0.97	0.78	0.45	0.88	0.91
6	(400,400)	0.61	0.97	0.96	0.55	0.90
7	(500,500)	0.77	0.92	0.86	0.64	0.89
8	(100,100,100)	1.00	0.70	0.12	0.98	0.88
9	(200,200,200)	0.98	0.76	0.34	0.91	0.91
10	(300,300,300)	0.94	0.78	0.44	0.77	0.87
11	(400,400,400)	0.81	0.91	0.84	0.69	0.90
12	(100,100,100,100)	0.92	0.83	0.60	0.79	0.90
13	(200,200,200,100)	0.93	0.83	0.62	0.80	0.89
14	(300,200,200,100)	0.67	0.96	0.94	0.58	0.90
15	(300,200,200,100,100)	0.95	0.83	0.60	0.85	0.91
16	(300,200,200,200,200)	0.82	0.88	0.76	0.67	0.87

(**Table 3.** Trial number alongside with trial's result)

4 Conclusion

In this paper, the MLPClassifier, a supervised learning type of Fully-Connected Neural Network was utilized to not only train a model on exoplanet classification, but to also determine which hyperparameters performed the best. The model was trained on NASA's Kepler data set with over 10000 candidates. The model with a performance of ≥ 0.9 AUC and ≥ 0.6 precision and recall rates had hyperparameters of (300,200,200,100,100) layers. With this accuracy, a NN model is proven useful for professional use, allowing more efficient processes in classifying candidates and detecting potential exoplanets, further helping scientists to meet their desired goal: to find an exoplanet similar to earth. Furthermore, data driven feature selection can be an alternate solution to improve the accuracy of machine learning models by selecting independent variables that are more important to the classification set, through a data driven selection that looks at its importance.

References

Arc. (2018, December 25). *Convolutional Neural Network. In this article, we will see what are… | by Arc*. Towards Data Science. Retrieved February 10, 2024, from https://towardsdatascience.com/convolutional-neural-network-17fb77e76c05

Brennan, P. (2019, November 19). *Why Do Scientists Search for Exoplanets? Here Are 7 Reasons*. Exoplanet Exploration. Retrieved February 3, 2024, from https://exoplanets.nasa.gov/news/1610/why-do-scientists-search-for-exoplanets-here-are-7-reasons/

DataBricks. (n.d.). *What is a Convolutional Layer?* Databricks. Retrieved February 10, 2024, from https://www.databricks.com/glossary/convolutional-layer

Gillis, A. S. (n.d.). *What is supervised learning? | Definition from TechTarget*. TechTarget. Retrieved February 9, 2024, from https://www.techtarget.com/searchenterpriseai/definition/supervised-learning

Hasan, F. (n.d.). *Educative Answers - Trusted Answers to Developer Questions*. Educative.io. Retrieved February 10, 2024, from https://www.educative.io/answers/what-are-some-deep-details-about-pooling-layers-in-cnn

IBM. (n.d.). *What is a Neural Network?* IBM. Retrieved March 3, 2024, from https://www.ibm.com/topics/neural-networks

Jin, Y., Yang, L., & Chiang, C.-E. (2022, April). IDENTIFYING EXOPLANETS WITH MACHINE LEARNING METHODS: A PRELIMINARY STUDY. *International Journal on Cybernetics & Informatics (IJCI)*, *11*(1/2), 32-42. 10.5121

Jock, N. (2023, June 21). *Convolutional Neural Network — Lesson 9: Activation Functions in CNNs*. Medium. Retrieved February 10, 2024, from https://medium.com/@nerdjock/convolutional-neural-network-lesson-9-activation-functions-in-cnns-57def9c6e759'

Kingma, Diederik P. and Jimmy Ba. “Adam: A Method for Stochastic Optimization.” CoRR abs/1412.6980 (2014): n. pag.

NASA. (2021, February 11). *Data columns in Kepler Objects of Interest Table*. NASA Exoplanet Archive. Retrieved March 3, 2024, from https://exoplanetarchive.ipac.caltech.edu/docs/API_kepcandidate_columns.html

NASA. (2021, April 2). *Overview | What is an Exoplanet? – Exoplanet Exploration: Planets Beyond our Solar System*. Exoplanet Exploration. Retrieved February 3, 2024, from https://exoplanets.nasa.gov/what-is-an-exoplanet/overview/

NASA & Bilogur, A. (2017, January 20). *Kepler Exoplanet Search Results*. Kaggle. Retrieved February 10, 2024, from https://www.kaggle.com/datasets/nasa/kepler-exoplanet-search-results

Park, Y.-S., & Lek, S. (2016). Chapter 7 - Artificial Neural Networks: Multilayer Perceptron for Ecological Modeling. In S. E. Jørgensen (Ed.), *Developments in Environmental Modelling* (Vol. 28, pp. 123-140). Elsevier. ISSN 0167-8892. ISBN 9780444636232. https://doi.org/10.1016/B978-0-444-63623-2.00007-4

(https://www.sciencedirect.com/science/article/pii/B9780444636232000074)

SciKit Learn. (n.d.). *1.17. Neural network models (supervised) — scikit-learn 1.4.1 documentation*. Scikit-learn. Retrieved March 3, 2024, from https://scikit-learn.org/stable/modules/neural_networks_supervised.html

SciKit Learn. (n.d.). *3.3. Metrics and scoring: quantifying the quality of predictions*. Scikit-learn. Retrieved March 3, 2024, from https://scikit-learn.org/stable/modules/model_evaluation.html#roc-metrics

Soni, M. (2020, October 7). *Convolution in Convolutional Neural Network(CNN) | by Manik Soni | Medium*. Manik Soni. Retrieved February 10, 2024, from https://maniksonituts.medium.com/convolution-in-convolutional-neural-network-cnn-53bf2a286427