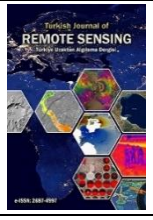
Araştırma Makalesi/Research Article
https://doi.org/10.51489/tuzal.1554248

2024; 6(2); 85-96

Turkish Journal of Remote Sensing

https://dergipark.org.tr/en/pub/tuzal

e-ISSN 2687-4997



EXOLIFE: Detection and Habitability Estimation of Exoplanets Using Machine Learning Techniques

Eren Yılmaz*[1], Muhammet Enes Artan[2], Ahmet Bilal Yanartaş[3]

[1] Hadımköy, Karaağaç Mah., 35000 Büyükçekmece, İstanbul
[2] Levent, Büyükdere Cd., 34394 Beşiktaş, İstanbul
[3] Acarlar Mah., 34800, Beykoz, İstanbul

Keywords
Exoplanet
NASA
Machine Learning
XGBoost
Biosignature

ABSTRACT

Exoplanets are among the most studied and remarkable topics in astronomy. Over the years, various methods have emerged for exoplanet detection, allowing for the identification of numerous exoplanet types. In this context, remote sensing and machine learning, which are central to our research, have significantly accelerated the detection process by leveraging algorithms. Our study involved training several machine learning models, including XGBoost, Random Forest, Multilayer Perceptron, K-Nearest Neighbor, Logistic Regression, and Support Vector Classifier, to compare their performance in both habitability assessment and exoplanet detection. The research utilized machine learning models trained on space observation data obtained from NASA, with the Python programming language serving as the foundation for the system's infrastructure. Our hypothesis was that "The detection of exoplanets and their evaluation within the scope of the habitability criterion can be increased to high accuracy rates with machine learning." Unlike merely detecting exoplanets, this study specifically aimed to identify Earth-like exoplanets. The XGBoost algorithm emerged as the most successful model in determining habitability, achieving an accuracy rate of 97.46% and demonstrating high precision and sensitivity. For exoplanet detection, all models achieved a main test accuracy rate of 96%; however, when considering sensitivity and precision, XGBoost was again the most effective. This research, following the synthesis and analysis of these two parameters, achieved a very high success rate compared to previous studies and made a significant contribution to the astronomy/astrophysics literature. Additionally, a Graphical User Interface (GUI) was developed, making the tested models functional through an application. The study successfully reached its goal of contributing important findings to the field.

EXOLIFE: Makine Öğrenmesi Kullanarak Ötegezegenlerin Tespit Edilmesi ve Yaşanabilirlik Tahmini Yapılması

Anahtar Kelimeler:
Ötegezegen
NASA
Makine Öğrenmesi
XGBoost
Biyoimza

ÖZ

Ötegezegenler, günümüzde astronomi alanında en çok çalışılan konular arasında yer almaktadır. Farklı türlerde oluşan ötegezegenlerin tespiti için çeşitli yöntemler geliştirilmiş ve bu sayede saptama mümkün hale gelmiştir. Bu araştırmada, ötegezegen tespiti için kullanılan uzaktan algılama ve makine öğrenmesi yöntemleri, algoritmalarla süreci hızlandırmaktadır. Projede, XGBoost, Rastgele Orman, Çok Katmanlı Algılayıcı, K-En Yakın Komşu, Lojistik Regresyon ve Destek Vektör Sınıflandırıcısı modelleri eğitilmiş ve hem yaşanılabilirlik hem de ötegezegen tespiti için karşılaştırmalar yapılmıştır. NASA verileriyle eğitilen bu makine öğrenmesi sistemi, Python yazılım diliyle oluşturulmuştur. Çalışma, "Ötegezegenlerin tespiti ve yaşanılabilirlik ölçütü kapsamında değerlendirilmesi makine öğrenmesi ile yüksek doğruluk oranlarına çıkarılabilir." hipotezine dayanarak Dünya benzeri ötegezegenleri bulmayı hedeflemiştir. Sonuçlarda, yaşanılabilirlik saptamasında %97.46 doğruluk oranı ile XGBoost algoritması en başarılı model olarak öne çıkmıştır. Gezegen tespitinde de %96'lık doğruluk oranıyla XGBoost, en başarılı model olmuştur. Araştırma, yüksek başarı oranıyla astronomi/astrofizik literatürüne önemli katkılar sağlamıştır. Ayrıca, çalışmanın sonucunda bir Grafiksel Kullanıcı Arayüzü (GUI) oluşturulmuş ve test edilen modeller işlevsel hale getirilmiştir.

Article Info

Received: 22/09/2024
Accepted: 28/11/2024
Published: 30/12/2024

Citation:

Yılmaz, E., Artan, M.E., Yanartaş, A. B. (2024). EXOLIFE: Detection and Habitability Estimation of Exoplanets Using Machine Learning Techniques. Turkish Journal of Remote Sensing, 6(2), 85-96.

***Corresponding Author:** Erem Yılmaz, ey0305739@gmail.com

1. INTRODUCTION

Exoplanets are planets that orbit stars outside the solar system. In recent years, scientists have carried out various studies on planets beyond the solar system and have planned long-term studies under this title (Patel, 2021). The idea of finding new habitable planets has been a topic of concern for researchers over the years, but as of today, exoplanets have made this issue a central issue again. It is predicted that exoplanets, which are also of great interest to the society, will become the center point of astronomy research in the coming years (Xin, 2022). When classifying exoplanets, scientists have divided them into four main categories based on their size: Terrestrial Planets, Super-Earths, Neptune-Like Planets, and Gas Giants (NASA, 2022). Terrestrial planets are rocky planets with Earth-like masses and iron-rich cores. Super-Earths, as the name suggests, are planets that are much larger than Earth but smaller than Uranus or Neptune. Neptune-like planets often have atmospheres with density H 2/He similar to those of Neptune (Helled, 2020). Finally, gas giants are gaseous planets with a size similar to or much larger than Saturn or Jupiter.

Various techniques have been developed for the detection of exoplanets and different areas such as remote sensing have become available through advancing technology. In this context, the four main known techniques are; the transition method is the radial velocity method, microlensing, and direct imaging (Dai, 2021). The transition method is currently the most useful technique. When a planet passes in front of a star, some of the starlight that is emitted bounces off that planet's atmosphere and reaches the earth. Through the transit method, as shown in Figure 1, the use of radiated light and gravitational force helps provide information about the planet's atmospheric chemical compositions and habitability. In the radial velocity method, velocity changes are used as determined by the changing direction of the gravitational force that any exoplanet receives from an outer planet as it rotates on the axis of another star (Huang, 2017). The microlensing method is a technique in which reproducible measurements cannot be made due to the fact that the measured event occurs very rarely, and therefore it is not used much. Finally, the direct imaging technique is a method that can detect exoplanets with an inclination of 90 degrees in their entirety, but this rarely happens due to the combination of the planet's small size and proximity to its star (Dai, 2021). Of the 5523 exoplanets confirmed by NASA so far, 1895 are marked as Neptune-like, 1748 as Gas Giants, 1674 as Super-Earths, 199 as Terrestrial, and the remaining 7 as unknown. Of the detected planets, 74.6% were determined using the transit method, 19.3% using the radial velocity method, 3.7% using the microlensing method, and the remaining 1.3% using the direct imaging method. (Brennan, 2020)

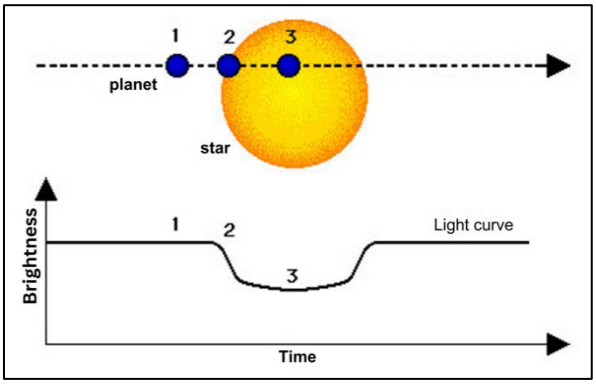


Figure 1. Transit Method

A biosignature is defined as a characteristic that provides scientific evidence for the existence of life and can be detected by remote sensing. These signatures are; gases in the atmosphere have several variations, such as chemical compounds and physical properties (Schwieterman, 2018). The habitable zone, on the other hand, is characterized when the conditions around a star can create a suitable environment for the existence of water. Since water is a critical component for life, determining the habitable zone is extremely important for assessing potential habitability (Ramirez, 2018). In this context, biological signature is processed under three basic subheadings. Among the biosignatures, gases in the atmosphere play a very important role. The presence of certain gases in the atmosphere, such as oxygen, methane, can be considered a biosignature when it is not in balance with the geology and chemistry of the planet. Among the gases in the Earth's atmosphere, there are gases such as N2, Ar, CO2, H2O, which are associated with living life and are directly related to biological activity (Yung, 2015). "Bioindicators" refer to the fact that atmospheric signatures can be produced by life or non-biotic processes. For example, although water (H_2O) is not a bioindicator, it is an important raw material and greenhouse gas for life. Other potential bioindicators include gases such as SO_2, H2S. These gases can be considered biosignatures when they are produced by volcanic activity or when they are present in the atmosphere under certain conditions (Meadows, 2018). The presence of plants and plant pigments on the planet's surface can be detected by a spectrum feature called the "Red Edge". This trait is unique to vegetation and indicates the presence of organisms that carry out photosynthesis (Seager, 2005). These observations were made with the aim of investigating potential biosignatures by remotely detecting the atmospheres and surfaces of Earth-like planets. The NASA Astrobiology Program has led efforts to search for biosignatures of exoplanet atmospheres. These studies were carried out by studying spectral models of the planet's surface and testing the concentration of biogenic gases (NASA, 2022).

Humanity's exoplanet exploration process has been going on for more than 30 years and is considered one of the most interesting branches of space exploration. At the very beginning of the process, in April 1984, the 2.5-metre du Pont telescope in Chile produced the original discovery image of the disc of dust and gas around the star Beta Pictoris (Hale, 2020). Following ongoing studies, the first exoplanets were discovered in January 1992, but they were unable to support organic life because they were bombarded with radiation from dead neutron stars in their orbits. In 1995, the first exoplanet orbiting a star similar to the Sun was discovered using the radial velocity technique. Subsequently, the use of technology became active; On April 4, 2001, the first planet in the "habitable zone" was found. Then, in October 2001, the first measurement of the atmosphere of a planet outside the solar system was made, and the first data were entered into the scientific world. As a result of subsequent studies, in 2005, the first detection of light from a planet outside the Solar System was made using the Spitzer Space Telescope. This event showed that Spitzer, designed to observe objects in the infrared spectrum, is a revolutionary tool in the characterization of exoplanets, and was an innovation that excited researchers (NASA, 2022). Then, in May 2007, the Spitzer Space Telescope was used to create the first map of an exoplanet's surface. In 2016, the Small Telescope for Transiting Planets and Planetesimals in Chile announced that it had found an exoplanet system containing at least seven planets. NASA's Kepler and K2 studies have discovered almost more than 2,600 new exoplanets using the transit method (Betz, 2023). Kepler played a crucial role in the discovery of these planets, many of which could be suitable places for life from outside our solar system. The James Webb Space Telescope, which is known as the most up-to-date and equipped telescope today and over which the most comprehensive studies have been carried out, has the ability to characterize the atmospheres of Earth-sized exoplanets as a large infrared telescope. One of the most important instruments that provide the telescope's qualities is the 'Large Binocular Telescope Interferometer', a NASA-funded instrument used to make high-resolution measurements and measure the absorption of dust orbiting stars (Brennan, 2021).

The discovery and characterization of extraterrestrial planets requires precise instrumentation and complex statistical methods. This process involves detecting weak planetary signals and modeling orbital and atmospheric features in detail. But the difficulties of sampling make it even more difficult to understand the characteristics of planetary populations derived from misleading or incomplete samples. The habitable zone does not only describe a certain distance at which a planet like Earth can be habitable, and it does not refer to the only location where habitable planets can exist. For example, the moons of giant planets in our own solar system can host habitable environments. However, while assessing the habitability of such regions in our own solar system is a difficult task, it is almost impossible to assess the habitability of similar environments in other star systems, and biosignatures in these specific regions may lose their perceptibility. However, with the advancement of technology, the detection of exoplanets by remote sensing systems has become a very popular method. Various studies conducted in this context have contributed to the literature in different fields. For example, in a study, he developed a new variational autoencoder algorithm to detect anomalies in exoplanet properties (Patel, 2023). This algorithm aimed to identify possible habitable exoplanets based on a broad set of features using unsupervised learning techniques. In another study, the proposed ASTRONET was carried out to analyze large and complex astronomical datasets using the deep learning architecture (Jagtap, 2021). Ishaani Priyadarshini and his team detect exoplanets by evaluating light intensity data using artificial intelligence and machine learning algorithms (Priyadarshini, 2021). Another study has developed an automated classification system to distinguish exoplanet transit signals using deep learning techniques (Mathur, 2020). Using data from the Kepler space telescope, Rajeev Mishra has developed a machine learning model that can classify exoplanets based on planet and star characteristics (Mishra, 2017). These studies show that machine learning and deep learning can provide great advantages when used in studies of the habitability of exoplanets. These techniques make it possible to process large and complex data sets produced by space telescopes and analyze properties associated with many planets and stars at the same time. These models help categorize planets as habitable or uninhabitable by detecting complex patterns and relationships. This supports the use of machine learning and deep learning as important tools for the search for habitability on exoplanets and increases our knowledge of potentially habitable worlds beyond our solar system.

2. METHOD

Within the scope of this research, research was first conducted using different machine learning models for the classification of habitable exoplanets. In addition, by analyzing the performance of each model, a comparative study was created on their efficiency (Kong et al., 2017). Although the study highlights both machine learning and deep learning as powerful tools, only machine learning techniques were applied in this research. Deep learning models were not employed, focusing instead on efficient, interpretable machine learning methods. The process followed in the research is shown in Figure 2. Accordingly, the system is divided into (1) database, (2) data preprocessing, (3) training

models, and (4) accuracy evaluation. Second, exoplanet detection was performed using different machine learning algorithms, as shown in Figure 3. The main programming language used in the research is Python. In addition, Pandas, Matplotlib, Scikit-Learn libraries were used.

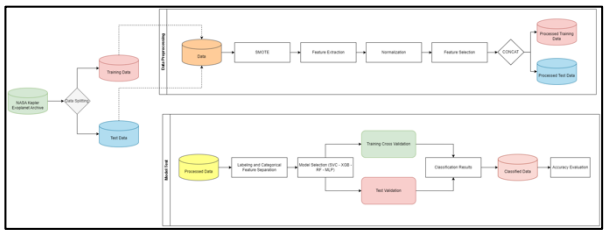


Figure 2. EXOLIFE exoplanet habitability model workflow diagram

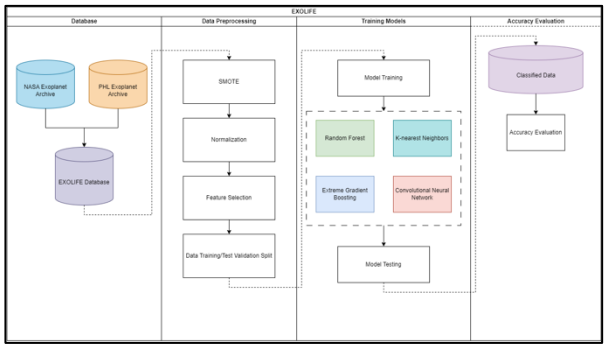


Figure 3. EXOLIFE exoplanet detection model workflow diagram.

2.1. The Data Collection

At the beginning of the research, data collection was carried out from different archives. In this context, data were collected from the NASA Exoplanet Archive, Kepler Mission Data and TESS (Transiting Exoplanet Survey Satellite) archives. While the NASA archive provides a wide range of data, the Kepler Mission and TESS data are particularly focused on observing transits (NASA, 2022). The dataset includes variables such as planetary mass, radius, distance from the host star, and atmospheric composition indicators where available. These data points are instrumental in detecting exoplanets and evaluating their potential habitability. Specific features—such as planetary mass and radius—aid in categorizing planets by type, while the orbital distance and stellar luminosity of host stars are critical for determining habitability zones.

2.2. Data Pre-Processing

It is very important to pre-process the data before introducing it to the models used. This process results in high-quality data or precise information, which has a direct impact on the model's ability to learn. Data pre-processing involved multiple stages to ensure high-quality inputs for model training. Initially, missing data points were managed using imputation. Outliers were identified and handled. Normalization was applied to rescale features to a 0-1 range, standardizing inputs across different measurement scales. Figure 4 shows the first version of the data set.

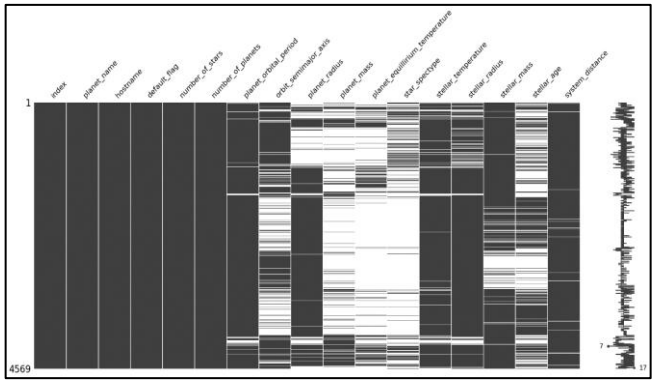


Figure 4. Raw data set

2.2.1. SMOTE

SMOTE is an acronym for 'Synthetic Minority Oversampling'. It is a method used to address data imbalances. This method is used to minimize dependency on majority-class values. As visualized in Figure 5, the data became more stable after this stage.

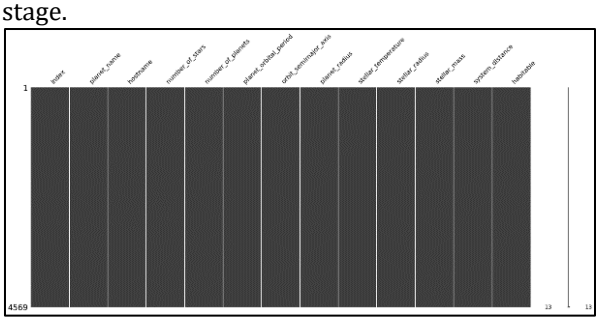


Figure 5. Post-SMOTE phase dataset

2.2.2. Normalization

The input dataset contains many features with different ranges, and normalization is helpful in bringing them to a similar scale. Values in the range [0, 1] are rescaled.

2.2.3. Feature selection

Feature selection is the process of identifying the most important and meaningful features in a data set (Mishra, 2017). This process ensures that the most appropriate features are selected for data analysis or machine learning models. The goal of feature selection is to improve model performance, reduce unnecessary or excess information, and prevent overfitting. In Figure 6, the mass-temperature relationship is visualized.

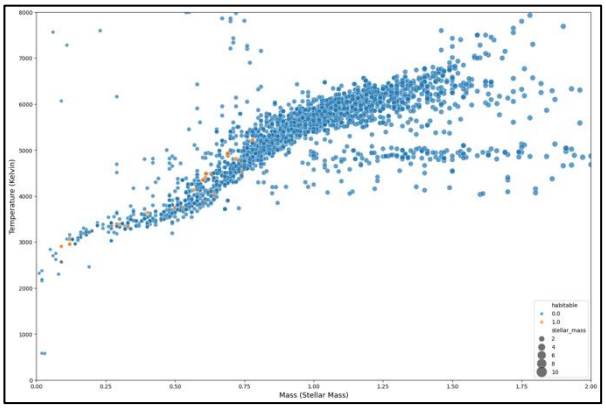


Figure 6. Temperature-mass relationship and habitability

2.2.4. Data clustering

Data clustering is a critical step that is often used in machine learning or data analysis processes. This step is done by dividing the dataset into three main sections: training, validation, and testing sets. The training set is used in the learning process of the model, while the validation set is used to evaluate the performance of the model and set hyperparameters. Finally, the test set is used to assess how well the model adapts to real-world data. By ensuring that the model is trained correctly and adapts to real data, data clustering helps us better understand the predictive ability of the model.

2.3. Models

2.3.1. Data exoplanet habitability prediction models

In this study, XGBoost, KNN, RF and LGR algorithms were used to predict the habitability of exoplanets (Jara-Maldonado et al., 2020).

XGBoost: It is a decision tree-based algorithm. This algorithm, which is based on the scikit-learn library, divides the data set into layers and makes optimal predictions. Based on the Gradient Boosting model, it minimizes errors and optimizes results.

K-Nearest Neighbor (KNN): It is a simple supervised learning algorithm. It makes predictions by placing nearby data points in the same class. Methods of calculating distances, such as Euclid or Minkowski, and optimization of the number of neighborhoods are important.

Random Forest (RF): It can be used for classification and regression purposes by combining various decision trees. For each data set, specific data is pulled and parameters are determined to improve the accuracy of the model.

Logistic Regression (LGR): It is an easy-to-apply classification method. With the Maximum Likelihood technique, a line is drawn separating the two classes, and this provides high accuracy rates overall.

2.3.2. Exoplanet detection models

In this research, the process of choosing among various classification models and the stages of training are detailed to predict the states of candidate planets and false positives. Model selection is based on the specific characteristics of each algorithm and the specific requirements of this study.

The Random Forest (RF) model represents an effective batch learning approach in capturing complex data relationships. This model has the potential to perform superiorly, especially when working with imbalanced datasets.

Support Vector Classification (SVC), on the other hand, offers a powerful alternative, especially for nonlinear classification tasks. SVC improves generalization by maximizing classification limits, which increases the stability and accuracy of the model.

Multilayer Perceptron (MLP) can effectively model nonlinear relationships using a deep learning structure. This multi-layered artificial neural network stands out for its ability to process complex data structures.

Finally, the XGBoost model both improves performance and has the capacity to deal with imbalanced datasets by using the gradient boosting technique.

By combining these models, it is aimed to increase the accuracy and reliability of predicting candidate planet states. This approach is intended to make significant contributions to the fields of astronomy and astrophysics.

2.4. Accuracy Assessment

The skicit-learn library used has allowed the work to be facilitated through various modules. The main modules and functions in the coding section are as follows:

- cross_val_score and KFold are both cross-validation methods.
- Metrics such as accuracy_score, balanced_accuracy_score, precision_score, recall_score, f1_score, and fbeta_score are used to measure the performance of classification models.
- make_scorer is used to create a custom score function.
- precision_recall_fscore_support function returns the classification report.
- The roc_curve, auc, and roc_auc_score functions are used for ROC curve analysis.
- The confusion_matrix function is used to evaluate the performance of a classification model.

These functions are taken as a basis when evaluating accuracy. While determining the extent to which the created software failed, the modules given

above were used and the result evaluation was carried out as a result of the results obtained from these modules.

In order to create a detailed confusion matrix, it is vital to apply 3 accuracy evaluation methods: sensitivity, precision and F-score. To begin with, sensitivity (D) was calculated by dividing the number of true positives within a class prediction by the total number of actual class instances, as shown in Eq. 1.

$$Sensitivity = \frac{TP}{TP+FN} \quad (1)$$

Precision (K), on the other hand, is applied by dividing the actual number of positive pixels by the total estimated number of pixels of a class, as seen in Eq. 2.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

The F measure (F1), which is used to evaluate Sensitivity and Precision in the same criterion, is used to provide the harmonic mean of them as given in Eq. 3. True Positive (DP) and False Negative (YN) variables used in sensitivity and precision calculations are values in the confusion metric that allow data sets to be tested for accuracy in different ways and show the accuracy of the classification made.

$$F1 = \frac{2 \times (D \times K)}{(D + K)} \quad (3)$$

2.5. GUI

The Graphical User Interface (GUI) developed within the scope of this research serves as an accessible platform that enables interaction with the machine learning model that performs the prediction of extraplanetary habitability. Built using Python's Tkinter library, the GUI provides an interface that includes input fields for off-planet parameters, a prediction trigger button, and an output screen that provides the model's habitability predictions. In the appendices section, there are interface images of the EXOLIFE application.

The GUI allows users to input planetary parameters and receive habitability predictions. This user-friendly interface is designed to make machine learning accessible for researchers and astronomers interested in real-time habitability assessments.

3. RESULTS & FINDINGS

3.1. Evaluation of the EXOLIFE Habitability Classification Model

So, the study results indicate a high overall accuracy of 97.46% for XGBoost in classifying exoplanets with potential habitability. Factors such as planetary size, mass, and orbital distance were critical in habitability estimation, supporting XGBoost's strength in handling diverse feature sets.

Table 1 shows the performance evaluation of the EXOLIFE habitability classification model using different machine learning algorithms. This evaluation includes various metrics that are used to understand the performance of the model. The XGBoost algorithm has the highest training accuracy rate. The model is able to learn the data with 99.32% accuracy during the training phase. At the same time, it has an accuracy rate of 97.46% during the testing phase. The sensitivity metric measures how accurate the model's positive predictions are, and the XGBoost model achieved a good result in this regard at 0.62%. In addition, metrics such as recall and F1 score are also high, indicating that the model provides a good balance of both true positive predictions and false positive predictions. Figure 7 shows the confusion matrix of the XGBoost model.

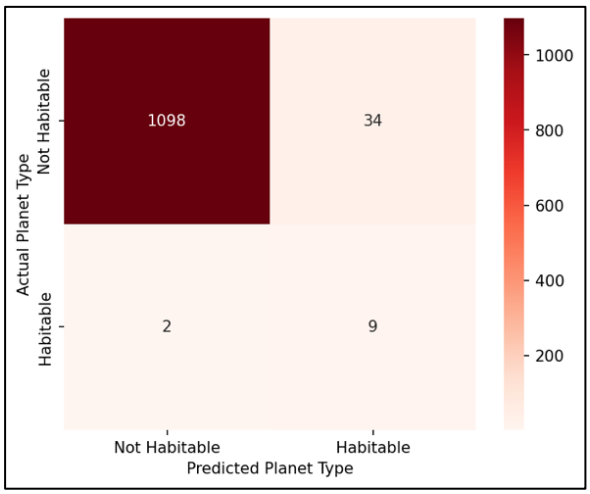


Figure 7. XGBoost confusion matrix

Looking at the results in Table 1, the KNN (K-Nearest Neighbor) algorithm also has high training and test accuracy rates (98.15% and 95.36%). However, the sensitivity and F1 score are lower, indicating that the model can make false positive predictions in some cases and has the potential to make improvements to the correct positive predictions.

The logistic regression model has lower training and test accuracy rates than the other two models (82.60% and 81.63%). The sensitivity and certainty values are moderate, indicating that there is potential to increase the model's accurate positive estimates.

The Random Forest algorithm achieved an excellent result in the training accuracy rate (100.00%) and the test accuracy rate is also high (98.34%) as given in Table 1. The sensitivity and precision values show a more balanced performance than other models, indicating that the model provides a good balance of true positive predictions and false positive predictions.

This high accuracy and sensitivity (0.90) suggest a significant correlation between these parameters and the habitability criteria established, confirming the efficacy of machine learning in exoplanet studies.

Table 1. Training and test results of habitability models

Model	Training Accuracy	Test Accuracy	Precision	Susceptibility	F1
XGBoost	99.32	97.46	0.62	0.90	0.69
KNN	98.15	95.36	0.57	0.84	0.49
Logistic Regression	82.60	81.63	0.52	0.82	0.49
Random Forest	100.00	98.34	0.62	0.68	0.64

3.2. Training Cross-Validation and Planet Detection Classification Model

The present study analyzes the performance of various classification models based on the results of instructional cross-validation. This analysis focuses specifically on the evaluation of Random Forest, XGBoost, Multilayer Perceptron (MLP), and Support Vector Classifier (SVC) models. These models were examined, especially in terms of 'sensitivity', 'precision', F1 score, and general accuracy parameters.

Random Forest (RF): This model demonstrated high sensitivity, recall, and an F1 score for both classes, achieving an overall accuracy of 96%, revealing that it had the capacity to effectively distinguish between "CANDIDATE" and "FALSE POSITIVE" samples.

XGBoost: In line with the Random Forest model, the XGBoost model also demonstrated consistent high sensitivity, precision, and an F1 score for both classes, demonstrating a strong performance with an accuracy rate of 96%.

Multilayer Sensor (MLP): This model was noted for its high sensitivity, precision, and F1 score for both classes, and was among the top-of-the-line models with an accuracy rate of 96%.

Support Vector Classifier (SVC): The SVC model achieved 96% accuracy, exhibiting high sensitivity, precision, and an F1 score for both classes.

As a result, the Random Forest, XGBoost, and Multilayer Detector models were identified as the best performing models with high sensitivity, precision, F1 score, and overall accuracy values, as shown in Figure 8. These findings suggest that these models are reliable options for classification tasks.

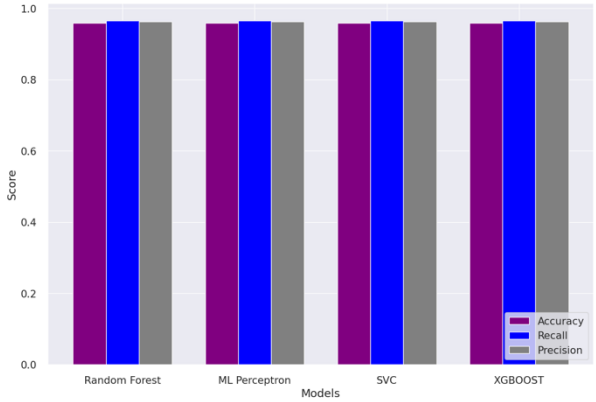


Figure 8. Training learning outcomes of detection models

3.3. Test Verification

This research on machine learning models evaluated the performance of the models, especially in the detection of candidate exoplanets. In the evaluation of the models, an analysis was made primarily on the sensitivity metric based on the research question. In this analysis, the XGBoost model, as shown in Table 2, stands out with a remarkable performance of 96.13%. This model showed a 96.50% sensitivity rate in detecting candidate exoplanets and a 95.64% success rate in identifying false positives, demonstrating its high competence in identifying true positive outcomes. In addition, with a precision of 96.62% and an accuracy of 95.48%, the XGBoost model has also shown an effective performance in minimizing false positive results. On the other hand, the Random Forest (RF) and Support Vector Classifier (SVC) models also achieved strong results, with high overall accuracy scores of 96.20% and 95.83%. These models have also shown impressive performance in accurately identifying true positives and reducing false positives.

Table 2. Exoplanet detection model results

Model	Test Accuracy	Susceptibility	Precision
Random Forest	0.962	[0.9559 0.9698]	[0.9761 0.9445]
MLP	0.955	[0.9507 0.9614]	[0.9695 0.9379]
SVC	0.958	[0.9520 0.9664]	[0.9734 0.9397]
XGBoost	0.959	[0.9649 0.9514]	[0.9624 0.9546]

In this context, based on the habitability criterion, the most successful study obtained after all data sets were trained and tested separately for each algorithm belonged to the XGBoost algorithm. The main test accuracy rate of 97.46% is seen as extremely high when studies on exoplanets are evaluated. In addition, the training accuracy rate of 99.32% obtained in the XGBoost algorithm appears to be a difficult rate to achieve. On the other hand, the XGBoost algorithm was well ahead of other algorithms in values such as sensitivity and F1 score, but it showed a result equivalent to the second most successful algorithm, Random Forest (RF), in terms of precision. RF, on the other hand, achieved 100% success by achieving an excellent rate in training accuracy, well above other compared algorithms. RF also stood out as the most successful algorithm in test accuracy rate with 98.34%. However, although the failure of the sensitivity score adversely affected the success of the RF algorithm, it remained constant at an average level in values such as precision and F1 score. On the other hand, KNN and Logistic Regression (LGR) algorithms were successful in a lower class than the other two algorithms. Although KNN did not perform poorly in the training set and main test accuracy rates, its failure in other metrics drew an incomplete image in the general scope. LGR, the most unsuccessful algorithm, was found to be around 80% in accuracy rates and was far away from the other three algorithms.

When the studies of S. Matheur et al. were evaluated, the Random Forest model reached 90%, the SVM model 88% and the KNN model 75%. This 2021 study 92 is seen as one of the leading current planetary detection studies. However, the accuracy rates obtained were well below the accuracy rate realized by us, and were even found to be in the same plane as the LGR model, which was described as the most unsuccessful. Therefore, the success shown in the study is seen at an advanced level considering the overseas studies and it seems that the research has successfully passed the livability determination accuracy rate test.

In another training, exoplanet detection was taken as a basis and the spectra of false positives and true positives were evaluated. In this context, the algorithms included in the study were RF, XGBoost, Multilayer Sensor and Support Vector Classifier. Although each of these has achieved an accuracy success rate of 96%, differences can be observed when examined in detail. The most successful algorithm at the overall level is the XGBoost algorithm. The XGBoost model seems to be quite successful with sensitivity and precision values of 0.95. The Random Forest algorithm, on the other hand, showed positive data with a sensitivity of 0.969 and a precision of 0.944. Although the SVC and MLP models do not seem to be low in general scope, they are below the other two algorithms, with a sensitivity of 0.96 and a precision of 0.93.

Table 3. Comparison of results with literature

Year and Author	Research	Method/Parameters	Results
Mislis et al. (2018)	Traversal of exoplanet light curves	Machine Learning Data Rejection Algorithm	Detection Efficiency ~ 80%
Zucker & Giryes (2018)	Detecting periodic transits of exoplanets	Deep Learning	Sensitivity = 0.94
Amin et al. (2018)	Detecting Exoplanet Systems	Adaptive Neuro-Fuzzy Systems	Accuracy~81%
Zingales & Waldmann (2018)	Reclaiming the Extraplanetary Atmosphere	Deep Convolutional Generative Adversarial Networks	300x speed increase over traditional buybacks
Ansdell et al. (2018)	Improvised Exoplanet Transit Classification	Deep Learning	2.0%–2.5% increase in model accuracy and average accuracy
Chintarungruangchai & Jiang (2019)	Detecting exoplanet transits	Machine Learning and CNN	Accuracy ~98%
Jara-Maldonado et al. (2020)	Research on Transiting Exoplanet Discovery	Machine Learning	Highest Accuracy achieved by Random Forests: 97.82
Sara Cuellar et al. (2022)	Exoplanet Detection with a Combination of Real and Synthetic Data	Deep Learning	Accuracy: 0.95

Exoplanet detection studies are followed closely at home and abroad and the number of researches is increasing. As can be seen in Table 3, the main studies that serve as examples for the general determination are in the 80% band, and the maximum rate is 98%, which is almost equivalent to the rate reached in this research. This clearly shows that the research is at a level that can compete with and even surpass the studies in the professional field and proves how comprehensive the research is.

Our research demonstrates an innovative approach that integrates machine learning approaches with traditional methods used in exoplanet data analysis. This is especially important in the context of processing and analyzing large data sets. The use of machine learning models speeds up the data analysis process and provides more accurate results. This allows for rapid and efficient progress in exoplanet research. In addition, the study shows the limitations of the methods used in exoplanet research and how machine learning

models can be applied to overcome these limitations. Analyses of the models' performance have demonstrated the potential of machine learning to overcome the challenges of exoplanet exploration. In particular, the use of these models to assess the habitability potential of exoplanets opens up new avenues for future research. Our research also takes an interdisciplinary approach to exoplanet science, offering a new perspective at the intersection of astrobiology, astronomy and computer science. Taking into account the complexity and multidisciplinary nature of exoplanet research, in particular, it contributes to the unification of knowledge and techniques in these areas. Transiting Exoplanet Survey Satellite (TESS) data were used in the data set of the study. TESS is a NASA-launched space mission that aims to scan most of the sky to detect thousands of new exoplanet candidates. The use of this data in our research significantly increases the innovative and up-to-date nature of the study.

The fact that the models used have many limitations has made it difficult to train the data sets throughout the process. Although the advantages of the models are more prominent when evaluated in general, some difficulties have prolonged the research process. XGBoost, for example, is a model that stands out for its high performance and fast training times, but it can be susceptible to a tendency to overfit. The Random Forest, on the other hand, is notable for its resistance to overlearning and feature importance, but it may require long training periods in large data sets. Logistic Regression can be effective in linear classification problems, but it can struggle to capture nonlinear relationships. KNN captures the local structure well and provides a clear model, but the cost of computation can increase with large data sets. SVC is useful in nonlinear classification problems, but large data sets may require long training times and correct parameter selection. MLP can learn complex relationships as a deep learning model, but it can be susceptible to overfitting, and hyperparameter tuning is important. Outliers and missing data should be managed in data collection and pre-processing processes. Model selection should be made depending on the characteristics of the data set, and methods such as cross-validation should be used in the accuracy evaluation process.

Apart from these, since exoplanets are a very current subject, it is very difficult to find a data set and reach the desired results. However, this can help to obtain more accurate results by increasing the data. Although the increase is possible with the addition of chemical data, the low number of scientists working on this subject today can be seen as a challenge. However, in this way, the quality of model training can be increased by using more data and accuracy rates can be kept constant in a successful plane.

The results of our study are important both scientifically and practically. Understanding the effectiveness of machine learning models in exoplanet classification and habitability assessment opens up new horizons in astrophysics and astrobiology. While the comprehensive analysis of the models contributes to the development of the methods used in exoplanet research, the applications of these models, especially on large data sets, allow data analysis processes to be accelerated and more accurate results to be obtained. Using data from innovative observational tools such as TESS provides an excellent opportunity to test the effectiveness and applicability of these models on real-world data.

4. CONCLUSION

This study investigates the use of various machine learning models to enhance the accuracy of exoplanet classification and predict habitability probabilities. The research compares the performance of different algorithms, including XGBoost, Logistic Regression, Random Forest, Multilayer Perceptron (MLP), and K-Nearest Neighbor (KNN), in the context of exoplanet classification. The models were evaluated on a variety of metrics such as training and test accuracy, sentiment, and F1 scores. This comparative analysis is crucial for understanding the potential and limitations of machine learning models in exoplanet research. Our research aims to contribute to the development of machine learning applications in exoplanet science, providing a solid foundation for further work in this area.

Evaluation of the findings obtained on the basis of different algorithms is very important in order to reach the most accurate and comprehensive result. Four different algorithms used in line with the method carried out drew separate conclusions from each other and facilitated concrete determinations in evaluating the processing of the data set specific to exoplanets.

In practical terms, this research helps to speed up and increase the efficiency of exoplanet exploration and assessment. In particular, the rapid and effective detection of habitable exoplanet candidates is critical for future space exploration and potentially humanity's efforts to colonize space. The findings of the research could provide important decision-making tools on issues such as space mission planning and prioritization of exoplanet observations. Furthermore, this interplay between machine learning and astrophysics could inspire innovative research in both fields and shape the direction of future scientific discoveries. The study can be considered as an important step in shaping the future of exoplanet science.

4.1 Recommendations

We suggest several ways to enhance the use of machine learning in discovering exoplanets. Firstly, employing a variety of machine learning algorithms can lead to more accurate and thorough classifications of exoplanets and assessments of their

habitability. While our study focused on models like XGBoost, Random Forest, Multilayer Perceptron (MLP), and K-Nearest Neighbor (KNN), we believe that more advanced techniques, such as deep learning, could offer additional benefits. Specifically, convolutional neural networks and recurrent neural networks could be particularly effective in analyzing visual and time-series data. These algorithms are likely to be key in better understanding and managing the complex and varied data associated with exoplanets.

Second, the inclusion of data from next-generation space telescopes such as the James Webb Space Telescope (JWST) will broaden the scope of the research and provide more detailed information. The high-resolution spectroscopic data provided by the JWST will allow for more detailed analysis of exoplanet atmospheres and offer new perspectives on habitability assessments. This data could allow machine learning models to make more precise and accurate predictions, helping usher in a new era of exoplanet research.

A third proposal is to estimate the chemical data on the atmospheric and surface compositions of exoplanets. Machine learning models can play a vital role in habitability analyses by analyzing spectroscopic data to predict the presence and concentrations of components in exoplanet atmospheres. This approach will contribute to a faster and more accurate detection of potentially habitable exoplanets. It is also possible for these models to simulate atmospheric and surface conditions to determine whether the conditions necessary for life exist.

Finally, considering that limited data sources on exoplanets pose a challenge, it is recommended to use data augmentation techniques. In addition to real data sets, synthetic data generation or diversification of existing data sets will allow machine learning models to be trained and tested on larger and more diverse data sets. This approach could provide significant benefits, especially in the classification and analysis of rare or little-known exoplanet species.

The implementation of these recommendations will maximize the potential of machine learning applications in exoplanet research and contribute to significant advances in this field. These developments will make valuable contributions to both the scientific community and humanity's effort to understand space.

Acknowledment

This study was supported by TUBITAK 2204-A Scientific Research Program.

Author contributions

E. Yılmaz: Literature review, conclusion and discussion sections, methodology (theoretical framework), contributions to astrophysics interpretations, and project administration.

M. E. Artan: Software development, findings analysis, methodology (algorithm implementation), and astrophysics-related evaluations.

A. B. Yanartaş: Introduction writing, compilation and organization of references, and contributions to the methodology (data collection and preprocessing).

Conflicts of Interest

The authors declare no conflict of interest.

Research and publication ethics statement

In the study, the author/s declare that there is no violation of research and publication ethics and that the study does not require ethics committee approval.

REFERENCES

Alei, E., Konrad, B. S., Angerhausen, D., Grenfell, J. L., Mollière, P., Quanz, S. P., ... & Wunderlich, F. (2022). Large Interferometer For Exoplanets (LIFE)-V. Diagnostic potential of a mid-infrared space interferometer for studying Earth analogs. Astronomy & Astrophysics, 665, A106. https://doi.org/10.1051/0004-6361/202243760

Angerhausen, D. (2019). Big Data and Machine Learning for Exoplanets and Astrobiology: Results from NASA Frontier Development Lab. In The Tenth Moscow Solar System Symposium (pp. 244-245). https://meetingorganizer.copernicus.org/EPSC-DPS2019/EPSC-DPS2019-588-1.pdf

Bapat, N. V., & Rajamani, S. (2023). Distinguishing Biotic vs. Abiotic Origins of 'Bio'signatures: Clues from Messy Prebiotic Chemistry for Detection of Life in the Universe. Life, 13(3), 766. https://doi.org/10.3390/life13030766

Basak, S., Saha, S., Mathur, A., Bora, K., Makhija, S., Safonova, M., & Agrawal, S. (2020). Ceesa meets machine learning: A constant elasticity earth similarity approach to habitability and classification of exoplanets. Astronomy and Computing, 30, 100335. https://doi.org/10.1016/j.ascom.2019.100335

Basant, R., Dietrich, J., & Apai, D. (2022). An Integrative Analysis of the Rich Planetary System of the Nearby Star e Eridani: Ideal Targets for Exoplanet Imaging and Biosignature Searches. The Astronomical Journal, 164(1), 12. https://doi.org/10.3847/1538-3881/ac6f58

Belenkaya, E. S., Alexeev, I. I., & Blokhina, M. S. (2022). Modeling of Magnetospheres of Terrestrial Exoplanets in the Habitable Zone around G-Type Stars. Universe, 8(4), 231. https://doi.org/10.3390/universe8040231

Claudi, R., & Alei, E. (2019). Biosignatures search in habitable planets. Galaxies, 7(4), 82. https://doi.org/10.3390/galaxies7040082

Cuéllar, S., Granados, P., Fabregas, E., Curé, M., Vargas, H., Dormido-Canto, S., & Farias, G. (2022). Deep learning exoplanets detection by combining real and synthetic data. Plos one, 17(5), e0268199. https://doi.org/10.1371/journal.pone.0268199

Dai, Z., Ni, D., Pan, L., & Zhu, Y. (2021, September). Five methods of exoplanet detection. In Journal of Physics: Conference Series (Vol. 2012, No. 1, p. 012135). IOP Publishing. https://doi.org/10.1088/1742-6596/2012/1/012135

Forestano, R. T., Matchev, K. T., Matcheva, K., & Unlu, E. B. (2023). Searching for Novel Chemistry in Exoplanetary Atmospheres Using Machine Learning for Anomaly Detection. The Astrophysical Journal, 958(2), 106. https://doi.org/10.3847/1538-4357/ad0047

Fujii, Y., Angerhausen, D., Deitrick, R., Domagal-Goldman, S., Grenfell, J. L., Hori, Y., ... & Stevenson, K. B. (2018). Exoplanet biosignatures: observational prospects. Astrobiology, 18(6), 739-778. https://doi.org/10.1089/ast.2017.1733

Hall, C., Stancil, P. C., Terry, J. P., & Ellison, C. K. (2023). A New Definition of Exoplanet Habitability: Introducing the Photosynthetic Habitable Zone. The Astrophysical Journal Letters, 948(2), L26. https://doi.org/10.48550/arXiv.2301.13836

Helled, R., Nettelmann, N., & Guillot, T. (2020). Uranus and Neptune: origin, evolution and internal structure. Space Science Reviews, 216, 1-26. https://doi.org/10.1007/s11214-020-00660-3

Huang, J. (2022). Planetary Science Meets Chemistry: Studying Potential Biosignature Gases in Terrestrial Exoplanet Atmospheres (Doctoral dissertation, Massachusetts Institute of Technology). https://dspace.mit.edu/bitstream/handle/1721.1/147547/Huang-huangjc-phd-chemistry-2022-thesis.pdf?sequence=1&isAllowed=y

Jagtap, R., Inamdar, U., Dere, S., Fatima, M., & Shardoor, N. B. (2021, April). Habitability of exoplanets using deep learning. In 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS) (pp. 1-6). IEEE. https://10.1109/IEMTRONICS52119.2021.9422571

Jara-Maldonado, M., Alarcon-Aquino, V., Rosas-Romero, R. *et al.* Transiting Exoplanet Discovery Using Machine Learning Techniques: A Survey. *Earth Sci Inform* **13**, 573–600 (2020). https://doi.org/10.1007/s12145-020-00464-7

Kaltenegger, L. (2017). How to characterize habitable worlds and signs of life. Annual Review of Astronomy and Astrophysics, 55, 433-485. https://doi.org/10.1146/annurev-astro-082214-122238

Kong, Z., Jiang, J. H., Burn, R., Fahy, K. A., & Zhu, Z. H. (2022). Analyzing the Habitable Zones of Circumbinary Planets Using Machine Learning. The Astrophysical Journal, 929(2), 187. https://doi.org/10.3847/1538-4357/ac5c5a

Krissansen-Totton, J., Thompson, M., Galloway, M. L., & Fortney, J. J. (2022). Understanding planetary context to enable life detection on exoplanets and test the Copernican principle. Nature Astronomy, 6(2), 189-198. https://doi.org/10.1038/s41550-021-01579-7

Meadows, V. S., Reinhard, C. T., Arney, G. N., Parenteau, M. N., Schwieterman, E. W., Domagal-Goldman, S. D., ... & Grenfell, J. L. (2018). Exoplanet biosignatures: understanding oxygen as a biosignature in the context of its environment. Astrobiology, 18(6), 630-662. https://doi.org/10.1089/ast.2017.1727

Mishra, R. (2017). Predicting habitable exoplanets from NASA's Kepler mission data using Machine Learning. *Predicting habitable exoplanets from NASA's Kepler mission data using Machine Learning*.

Novak, R., Bradak, B., Kovacs, J., & Gomez, C. (2023). Search for Exoplanets with a Possible Surface Water Ocean. Physical Sciences Forum, 7(1), 19. https://doi.org/10.3390/ECU2023-14020

Priyadarshini, I., Puri, V. A convolutional neural network (CNN) based ensemble model for exoplanet detection. Earth Sci Inform 14, 735–747 (2021). https://doi.org/10.1007/s12145-021-00579-5

Ramirez, R. M. (2018). A More Comprehensive Habitable Zone for Finding Life on Other Planets. Geosciences, 8(8), 280. https://doi.org/10.3390/geosciences8080280

Ranjan, S., Seager, S., Zhan, Z., Koll, D. D., Bains, W., Petkowski, J. J., ... & Lin, Z. (2022). Photochemical Runaway in Exoplanet

Atmospheres: Implications for Biosignatures. The Astrophysical Journal, 930(2), 131. https://doi.org /10.3847/1538-4357/ac5749

Schwieterman EW, Kiang NY, Parenteau MN, Harman CE, DasSarma S, Fisher TM, Arney GN, Hartnett HE, Reinhard CT, Olson SL, Meadows VS, Cockell CS, Walker SI, Grenfell JL, Hegde S, Rugheimer S, Hu R, Lyons TW. Exoplanet Biosignatures: A Review of Remotely Detectable Signs of Life. Astrobiology. 2018 Jun;18(6):663-708. https://doi.org/ 10.1089/ast.2017.1729.

Seager, S. (2014). The future of spectroscopic life detection on exoplanets. Proceedings of the National Academy of Sciences, 111(35), 12634-12640. https://doi.org/10.1073/pnas.1304213111

Seager, S., & Bains, W. (2015). The search for signs of life on exoplanets at the interface of chemistry and planetary science. Science advances, 1(2), e1500047. https://doi.org/ 10.1126/sciadv.1500047

Soboczenski, F., Himes, M. D., O'Beirne, M. D., Zorzan, S., Baydin, A. G., Cobb, A. D., ... & Domagal-Goldman, S. D. (2018). Bayesian deep learning for exoplanet atmospheric retrieval. arXiv preprint arXiv:1811.03390. https://doi.org/10.48550/arXiv.1811.03390

Thompson, M. A., Krissansen-Totton, J., Wogan, N., Telus, M., & Fortney, J. J. (2022). The case and context for atmospheric methane as an exoplanet biosignature. Proceedings of the National Academy of Sciences, 119(14), e2117933119. https://doi.org/10.1073/pnas.2117933119

Tuchow, N. W., & Wright, J. T. (2020). A Framework for Relative Biosignature Yields from Future Direct Imaging Missions. The Astrophysical Journal, 905(2), 108. https://doi.org/10.3847/1538-4357/abc556

Xin, L. (2022). Exoplanets, extraterrestrial life and beyond: an interview with Douglas Lin. National Science Review, 9(2), nwac008. https://doi.org/10.1093/nsr/nwac008