Exoplanetary Surface Composition Prediction Using Machine Learning

Dibya Bharati Pradhan[1] **Oommen P Jose**[1]
[1]School of Physical Sciences
National Institute of Science Education and Research, HBNI
Jatni-752050, India
dibyabharati.pradhan@niser.ac.in
oommen.jose@niser.ac.in

Abstract

The search for extraterrestrial life has been a major goal of space exploration for decades. The new generation of telescopes, such as ELTs and the JWST, is designed to obtain the measurements of the atmospheric composition of Earth-sized exoplanets. However, obtaining these measurements through spectroscopy is time-consuming, even for these advanced telescopes. So, Photometry has been proposed as a promising way to rapidly classify and prioritize exoplanets. In this report, we explore the feasibility of Machine Learning to predict the composition of the surface of Earth-like terrestrial exoplanets by analyzing the Photometric flux data that is anticipated from advanced telescopes in the future. Our approach involves employing a combination of Atmos, PSG, and PICASO models to generate the data set and running the SVR, Random Forest and Neural Network Algorithms to achieve the objective.

1 Introduction

Till date, above 5000 exoplanets have been detected and a few dozen of them have been found in the habitable zone. It is a region around a star where liquid water can exist on the surface of the planet. To determine whether a planet is potentially habitable, its atmosphere and surface properties need to be studied in detail. One way to assess the atmosphere of an exoplanet is to observe its spectrum, looking for specific biosignatures, such as the presence of oxygen or methane, which could indicate the presence of life. But even if such biosignatures are detected, they may not necessarily be proof of life, as there could be other natural processes that produce them. Another approach to searching for life on exoplanets is to look for surface features that could indicate the presence of living organisms. Several studies have shown that photometric colors of planetary bodies can distinguish between icy, rocky, and gaseous surface types. They have also showed that models of Earth-like exoplanets lie in a certain color space. Hence we focus on probing the exoplanetary surface.

2 Exoplanetary surface probing using Machine Learning

Previous works (Pham & Kaltenegger, 2022, 2021) in the field have investigated the feasibility of using broad-band filter photometry combined with machine learning and Markov chain Monte Carlo (MCMC) to detect water on the surface of earth like exoplanets in various forms like snow, clouds, and liquid water. They trained the XGBoost, a machine-learning algorithm to perform binary classification and to predict on the presence of snow, clouds, and water on the exoplanet's surface using photometry. For snow and cloud, the algorithm achieved a high balanced accuracy (90%) for a given S/N ≥ 20. It predicted up to 70% balanced accuracy for liquid water. The paper also identified

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

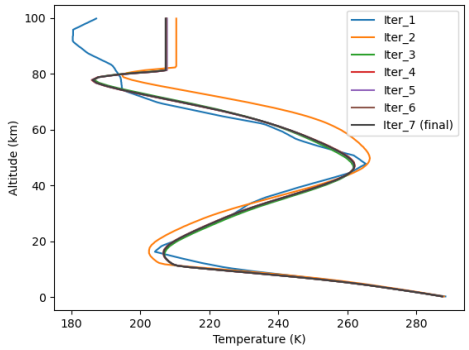

Figure 1: Pressure (altitude) Temperature profile

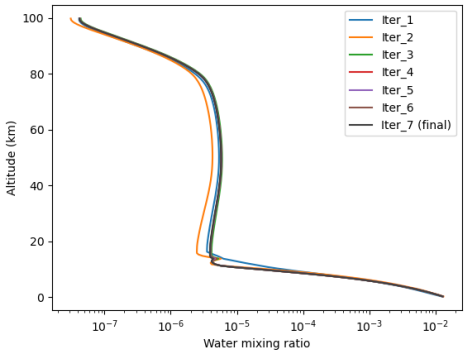

Figure 2: Altitude vs Water Mixing Ratio

five optimal filters to identify snow, clouds, and liquid water on a terrestrial planet's surface based on XGBoost's feature importance ranking, which could be implemented in telescope designs that search for water on exoplanets.

To test these optimal filters, Bayesian inferences using MCMC are performed on an Earth case study using around 100 random realizations (of the planetary models). For finding snow and clouds, the results showed optimal accuracy for most predictions at S/N $\geq$ 50 within about 5% of the true flux. Liquid water detection was more challenging, but most predictions were within 20% of the true values.

The reflection spectra for Cold Earth-like Exoplanets with modern outgassing rates, the surface temperature of 273K and a reduced solar spectrum luminosity of 0.875 $L_\odot$ are generated using an advanced coupled 1D Photochemical Climate model, Exo-Prime2 (Madden & Kaltenegger, 2020), in conjunction with a Radiative Transfer component. The model incorporates molecules like $C_2H_6, CH_4, CO, CO_2, H_2CO, H_2O, etc$, which are critical for spectroscopic analyses.

To account for surface reflectivity, the study selects six major components, namely water, snow, basalt, vegetation (aspen leaf), sand, and clouds, whose albedos are integrated into the model.The albedo of the cloud is obtained from the MODIS 20 μm cloud model (King et al., 1997), whereas the albedo of other major surface components is obtained from the USGS Spectra Library.

The atmospheric profile is held constant for varying combinations of reflective components. The model implements nine ideal broad-band filters (f_i) with a width of 0.2 μm each within the wavelength range to generate corresponding true flux values through integration. Gaussian noise is added to these values to create the observed flux values that serve as the data set for the machine learning algorithm.

Similarly, (Pham & Kaltenegger, 2021) has investigated the identificaion of surface features and biota classification using photometry with Johnson filters which can prioritize exoplanets for further study. They created a reflection spectra grid for terrestrial earth-like planets with varying surface compositions and cloud coverage. It assesses the sensitivity of the results to six different biota samples, which includes vegetation and UV resistant biota. The balanced accuracy varied between 50% and 75% for different algorithms. It also depended on the signal-to-noise ratio.

We obtained the Albedo values from the Planetary Spectrum Generator (PSG) which is given as input to PICASO along with PT profile and Abundance files generated from Atmos. PICASO is equipped with a surface reflectivity module which we will use to generate the Reflection Spectra. The albedo values from PSG closely matched the values utilized in the reference paper.

3 Baseline Algorithms

In (Pham & Kaltenegger, 2022), the authors use XGBoost algorithm on simulated photometric data to identify characteristic features of water, snow, and clouds. The authors test their method on a set of synthetic exoplanet spectra and show that their approach is effective at detecting the presence of

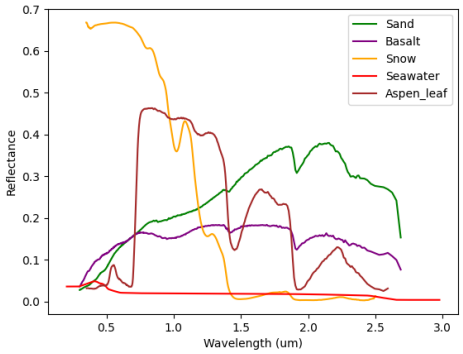

Figure 3: Albedos of components, given as input to PICASO

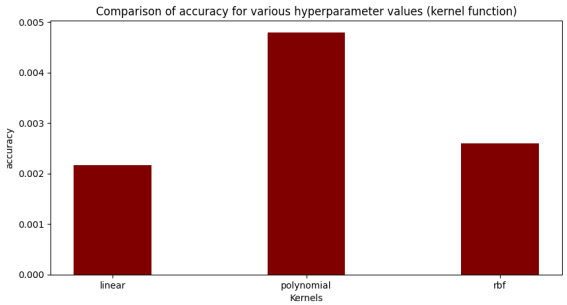

Figure 4: Comparison of accuracy for different kernel functions in SVR

water, snow, and clouds even in the presence of noise and other sources of uncertainty. They vary the signal-to-noise ratios and compare the accuracy values.

In (Pham & Kaltenegger, 2021), the authors compare the seven different algorithms based on how effective they are at predicting the presence of a certain kind of biota on the surface of exoplanets from simulated photometric flux with noise added. The algorithms compared in the paper are,

Linear Discriminant Analysis (LDA) K-Nearest Neighbors (KNN), Classification and Regression Tree (CART), Logistic Regression (LR), Naive-Bayes Classifier (NB), Support Vector Machine (SVM), Random Forest (RF), Majority voting classifier, Hard Majority Voting(HMV) and Soft Majority Voting(SMV). The signal-to-noise ratio is varied in this case and the accuracy values of each of these classification algorithms are compared in the paper.

4 Experiment

Github link Our codes are contained in the following link.

In this project, we aim to replicate (Pham & Kaltenegger, 2022)'s technique by using two distinct models (Atmos and PICASO) instead of the conventional model, EXOPRIME2. We are also looking at the feasibility of ML in predicting the exact surface composition of exoplanets by training the algorithms first with the available data set in (Pham & Kaltenegger, 2022) and then training them on the data generated by our models. The Atmos model we used is the Photochemical Climate model (Arney et al., 2016), much like EXOPRIME2, which we coupled with the Radiative Transfer model PICASO (Batalha et al., 2019). Atmos runs iteratively till it reaches convergence. A plot of Atmos finding convergence for the P-T profiles Fig:1, Water mixing ratio Fig:2 are shown.
We are trying to predict the percentage surface composition of exoplanets from simulated photometric flux data using regression algorithms instead of classification. We have implemented Support Vector Regression (SVR) as well as Random forest Regression to predict the surface composition from data without noise and compared the accuracy of SVR for various kernel functions. The hyperparameters in the case of SVR are the kernel function while that for RFR is the number of trees considered. The plot of the comparison is shown above Fig:4. We can see that out of the three the linear kernel performs best for this dataset. The MSE error obtained without adding noise to the dataset for the SVR using linear kernel for nine filters is 0.00216765 while for RFR with the number of trees taken to be 100 is 0.0035305. The balanced accuracy obtained for classification in the first paper for 100 signal-to-noise ratio for XGBoost was above 95 percent.

The dataset comprises reflective photometric flux data obtained by applying filters to the reflection spectra of different surface combinations, including Sand, Sea water, Vegetation, Cloud, Basalt, and Snow. The dataset covers all possible percentage combinations of these six surfaces with a step size of 5 percent as shown in the table below,

Table 1: Surface combinations (Labels)

	cloud	snow	sand	seawater	basalt	veg
0	0.00	0.00	0.00	0.00	0.00	1.00
1	0.00	0.00	0.00	0.00	0.05	0.95
...	...	...	...	...	...	...
53128	0.95	0.05	0.00	0.00	0.00	0.00
53129	1.00	0.00	0.00	0.00	0.00	0.00

resulting in a total of 53,130 combinations over a wavelength range spanning from 0.41 μm to 2.35 μm. These reflection spectra were obtained from the previously mentioned in the first paper. , and the photometric flux values were derived by applying filters to the reflection spectra. A model was then developed to predict the percentage surface combination of the six surfaces considered based on the photometric flux values. Here the features are the flux values obtained after applying the filters. A sample of features is shown in Table 2. The number of filters considered is varied and the accuracy of the model is compared in the plot below.

Table 2: Flux values for 9 filters (Features)

	f1	f2	f3	f4	f5	f6	f7	f8	f9
0	66.4502	70.5596	60.5511	34.8012	13.2305	8.8234	6.3379	0.3508	1.5993
1	67.5648	69.2096	58.6436	33.7876	12.8987	8.7628	6.2505	0.3810	1.6354
...	...	...	...	...	...	...	...	...	...
53129	241.7693	147.7095	93.5636	60.7461	34.6885	22.7134	17.6443	4.5057	6.1786

The labels taken for the model are the surface combinations as shown in Table 1 **??**

5 Further Plan

We would obtain Albedo for the cloud from (King et al., 1997). Then we plan to compare the spectra obtained by our method and the spectra given in the data set of the paper (Pham & Kaltenegger, 2022) for a particular surface combination. This will ensure that our method is able to reproduce the paper's data set. We will then use our created spectra to train the algorithms that we already ran on the data set of the paper. We will split our data set randomly into 80:20 ratios to serve as our training and validation set. The validation set will then be augmented with Gaussian noise and accuracy will be checked. Last, we plan to predict the composition upon training the data using MLP Neural Network. Further, we also wish to find optimal filters among the nine optimal filters used in the model which will help future telescopes in prioritizing the targets of Exoplanets for spectroscopic analysis.

References

Arney G., et al., 2016, Astrobiology, 16, 873

Batalha N. E., Marley M. S., Lewis N. K., Fortney J. J., 2019, The Astrophysical Journal, 878, 70

King M., Tsay S.-C., Platnick S., Wang M., Liou K., 1997, doi:https: //modis.gsfc.nasa.gov/data/atbd/atbd mod05.pdf

Madden J., Kaltenegger L., 2020, Monthly Notices of the Royal Astronomical Society, 495, 1

Pham D., Kaltenegger L., 2021, Monthly Notices of the Royal Astronomical Society, 504, 6106

Pham D., Kaltenegger L., 2022, Monthly Notices of the Royal Astronomical Society: Letters, 513, L72

Dibya Bharati Pradhan Paper Check

By Dibya Bharati Pradhan

WORD COUNT	1985	TIME SUBMITTED	10-MAR-2023 09:17PM
		PAPER ID	97476620

Exoplanetary Surface Composition Prediction Using Machine Learning

Dibya Bharati Pradhan[1] **Oommen P Jose**[1]
[1]School of Physical Sciences
National Institute of Science Education and Research, HBNI
Jatni-752050, India
dibyabharati.pradhan@niser.ac.in
oommen.jose@niser.ac.in

Abstract

The search for extraterrestrial life has been a major goal of space exploration for decades. The new generation of telescopes, such as ELTs and the JWST, is designed to obtain the measurements of the atmospheric composition of Earth-sized exoplanets. However, obtaining these measurements through spectroscopy is time-consuming, even for these advanced telescopes. So, Photometry has been proposed as a promising way to rapidly classify and prioritize exoplanets. In this report, we explore the feasibility of Machine Learning to predict the composition of the surface of Earth-like terrestrial exoplanets by analyzing the Photometric flux data that is anticipated from advanced telescopes in the future. Our approach involves employing a combination of Atmos, PSG, and PICASO models to generate the data set and running the SVR, Random Forest and Neural Network Algorithms to achieve the objective.

1 Introduction

Till date, above 5000 exoplanets have been detected and a few dozen of them have been found in the habitable zone. It is a region around a star where liquid water can exist on the surface of the planet. To determine whether a planet is potentially habitable, its atmosphere and surface properties need to be studied in detail. One way to assess the atmosphere of an exoplanet is to observe its spectrum, looking for specific biosignatures, such as the presence of oxygen or methane, which could indicate the presence of life. But even if such biosignatures are detected, they may not necessarily be proof of life, as there could be other natural processes that produce them. Another approach to searching for life on exoplanets is to look for surface features that could indicate the presence of living organisms. Several studies have shown that photometric colors of planetary bodies can distinguish between icy, rocky, and gaseous surface types. They have also showed that models of Earth-like exoplanets lie in a certain color space. Hence we focus on probing the exoplanetary surface.

2 Exoplanetary surface probing using Machine Learning

Previous works (Pham & Kaltenegger, 2022, 2021) in the field have investigated the feasibility of using broad-band filter photometry combined with machine learning and Markov chain Monte Carlo (MCMC) to detect water on the surface of earth-like exoplanets in various forms like snow, clouds, and liquid water. They trained the XGBoost, a machine-learning algorithm to perform binary classification and to predict on the presence of snow, clouds, and water on the exoplanet's surface using photometry. For snow and cloud, the algorithm achieved a high balanced accuracy (90%) for a given S/N $\geq$ 20. It predicted up to 70% balanced accuracy for liquid water. The paper also identified

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

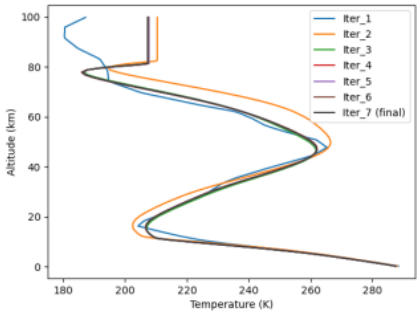

Figure 1: Pressure (altitude) Temperature profile

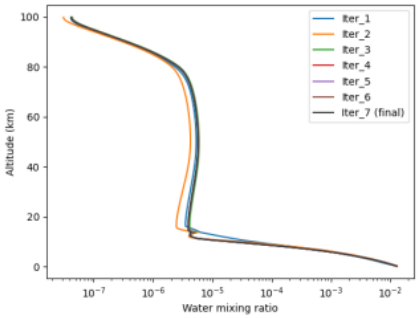

Figure 2: Altitude vs Water Mixing Ratio

five optimal filters to identify snow, clouds, and liquid water on a terrestrial planet's surface based on XGBoost's feature importance ranking, which could be implemented in telescope designs that search for water on exoplanets.

To test these optimal filters, Bayesian inferences using MCMC are performed on an Earth case study using around 100 random realizations (of the planetary models). For finding snow and clouds, the results showed optimal accuracy for most predictions at S/N $\geq$ 50 within about 5% of the true flux. Liquid water detection was more challenging, but most predictions were within 20% of the true values.

The reflection spectra for Cold Earth-like Exoplanets with modern outgassing rates, the surface temperature of 273K and a reduced solar spectrum luminosity of 0.875 $L_{\odot}$ are generated using an advanced coupled 1D Photochemical Climate model, Exo-Prime2 (Madden & Kaltenegger, 2020), in conjunction with a Radiative Transfer component. The model incorporates molecules like $C_2H_6, CH_4, CO, CO_2, H_2CO, H_2O, etc$, which are critical for spectroscopic analyses.

To account for surface reflectivity, the study selects six major components, namely water, snow, basalt, vegetation (aspen leaf), sand and clouds, whose albedos are integrated into the model.The albedo of the cloud is obtained from the MODIS 20 μm cloud model (King et al., 1997), whereas the albedo of other major surface components is obtained from the USGS Spectra Library.

The atmospheric profile is held constant for varying combinations of reflective components. The model implements nine ideal broad-band filters (f_i) with a width of 0.2 μm each within the wavelength range to generate corresponding true flux values through integration. Gaussian noise is added to these values to create the observed flux values that serve as the data set for the machine learning algorithm.

Similarly, (Pham & Kaltenegger, 2021) has investigated the identificaion of surface features and biota classification using photometry with Johnson filters which can prioritize exoplanets for further study. They created a reflection spectra grid for terrestrial earth-like planets with varying surface compositions and cloud coverage. It assesses the sensitivity of the results to six different biota samples, which includes vegetation and UV resistant biota. The balanced accuracy varied between 50% and 75% for different algorithms. It also depended on the signal-to-noise ratio.

We obtained the Albedo values from the Planetary Spectrum Generator (PSG) which is given as input to PICASO along with PT profile and Abundance files generated from Atmos. PICASO is equipped with a surface reflectivity module which we will use to generate the Reflection Spectra. The albedo values from PSG closely matched the values utilized in the reference paper.

3 Baseline Algorithms

In (Pham & Kaltenegger, 2022), the authors use XGBoost algorithm on simulated photometric data to identify characteristic features of water, snow, and clouds. The authors test their method on a set of synthetic exoplanet spectra and show that their approach is effective at detecting the presence of

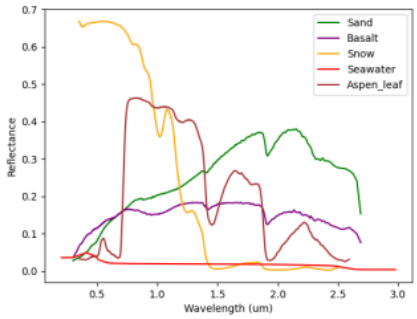

Figure 3: Albedos of components, given as input to PICASO

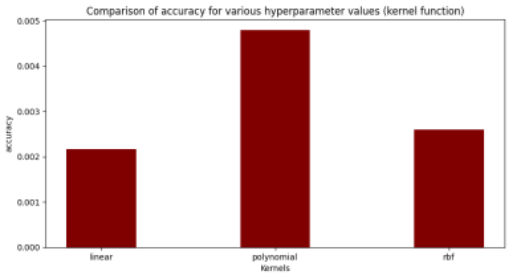

Figure 4: Comparison of accuracy for different kernel functions in SVR

water, snow, and clouds even in the presence of noise and other sources of uncertainty. They vary the signal-to-noise ratios and compare the accuracy values.

In (Pham & Kaltenegger, 2021), the authors compare the seven different algorithms based on how effective they are at predicting the presence of a certain kind of biota on the surface of exoplanets from simulated photometric flux with noise added. The algorithms compared in the paper are,

Linear Discriminant Analysis (LDA) K-Nearest Neighbors (KNN), Classification and Regression Tree (CART), Logistic Regression (LR), Naive-Bayes Classifier (NB), Support Vector Machine (SVM), Random Forest (RF), Majority voting classifier, Hard Majority Voting(HMV) and Soft Majority Voting(SMV). The signal-to-noise ratio is varied in this case and the accuracy values of each of these classification algorithms are compared in the paper.

4 Experiment

Github link Our codes are contained in the following link

In this project, we aim to replicate (Pham & Kaltenegger, 2022)'s technique by using two distinct models (Atmos and PICASO) instead of the conventional model, EXOPRIME2. We are also looking at the feasibility of ML in predicting the exact surface composition of exoplanets by training the algorithms first with the available data set in (Pham & Kaltenegger, 2022) and then training them on the data generated by our models. The Atmos model we used is the Photochemical Climate model (Arney et al., 2016), much like EXOPRIME2, which we coupled with the Radiative Transfer model PICASO (Batalha et al., 2019). Atmos runs iteratively till it reaches convergence. A plot of Atmos finding convergence for the P-T profiles Fig 1, Water mixing ratio Fig 2 are shown.
We are trying to predict the percentage surface composition of exoplanets from simulated photometric flux data using regression algorithms instead of classification. We have implemented Support Vector Regression (SVR) as well as Random forest Regression to predict the surface composition from data without noise and compared the accuracy of SVR for various kernel functions. The hyperparameters in the case of SVR are the kernel function while that for RFR is the number of trees considered. The plot of the comparison is shown above Fig 4. We can see that out of the three the linear kernel performs best for this dataset. The MSE error obtained without adding noise to the dataset for the SVR using linear kernel for nine filters is 0.00216765 while for RFR with the number of trees taken to be 100 is 0.0035305. The balanced accuracy obtained for classification in the first paper for 100 signal-to-noise ratio for XGBoost was above 95 percent.

The dataset comprises reflective photometric flux data obtained by applying filters to the reflection spectra of different surface combinations, including Sand, Sea water, Vegetation, Cloud, Basalt, and Snow. The dataset covers all possible percentage combinations of these six surfaces with a step size of 5 percent as shown in the table below,

Table 1: Surface combinations (Labels)

	cloud	snow	sand	seawater	basalt	veg
0	0.00	0.00	0.00	0.00	0.00	1.00
1	0.00	0.00	0.00	0.00	0.05	0.95
...	...	...	...	...	...	...
53128	0.95	0.05	0.00	0.00	0.00	0.00
53129	1.00	0.00	0.00	0.00	0.00	0.00

resulting in a total of 53,130 combinations over a wavelength range spanning from 0.41 μm to 2.35 μm. These reflection spectra were obtained from the previously mentioned in the first paper. , and the photometric flux values were derived by applying filters to the reflection spectra. A model was then developed to predict the percentage surface combination of the six surfaces considered based on the photometric flux values. Here the features are the flux values obtained after applying the filters. A sample of features is shown in Table 2. The number of filters considered is varied and the accuracy of the model is compared in the plot below.

Table 2: Flux values for 9 filters (Features)

	f1	f2	f3	f4	f5	f6	f7	f8	f9
0	66.4502	70.5596	60.5511	34.8012	13.2305	8.8234	6.3379	0.3508	1.5993
1	67.5648	69.2096	58.6436	33.7876	12.8987	8.7628	6.2505	0.3810	1.6354
...	...	...	...	...	...	...	...	...	...
53129	241.7693	147.7095	93.5636	60.7461	34.6885	22.7134	17.6443	4.5057	6.1786

The labels taken for the model are the surface combinations as shown in Table 1 ??

5 Further Plan

We would obtain Albedo for the cloud from (King et al., 1997). Then we plan to compare the spectra obtained by our method and the spectra given in the data set of the paper (Pham & Kaltenegger, 2022) for a particular surface combination. This will ensure that our method is able to reproduce the paper's data set. We will then use our created spectra to train the algorithms that we already ran on the data set of the paper. We will split our data set randomly into 80:20 ratios to serve as our training and validation set. The validation set will then be augmented with Gaussian noise and accuracy will be checked. Last, we plan to predict the composition upon training the data using MLP Neural Network. Further, we also wish to find optimal filters among the nine optimal filters used in the model which will help future telescopes in prioritizing the targets of Exoplanets for spectroscopic analysis.

References

Arney G., et al., 2016, Astrobiology, 16, 873

Batalha N. E., Marley M. S., Lewis N. K., Fortney J. J., 2019, The Astrophysical Journal, 878, 70

King M., Tsay S.-C., Platnick S., Wang M., Liou K., 1997, doi:https: //modis.gsfc.nasa.gov/data/atbd/atbd_mod05.pdf

Madden J., Kaltenegger L., 2020, Monthly Notices of the Royal Astronomical Society, 495, 1

Pham D., Kaltenegger L., 2021, Monthly Notices of the Royal Astronomical Society, 504, 6106

Pham D., Kaltenegger L., 2022, Monthly Notices of the Royal Astronomical Society: Letters, 513, L72

Dibya Bharati Pradhan Paper Check

ORIGINALITY REPORT

18%
SIMILARITY INDEX

PRIMARY SOURCES

1	academic.oup.com Internet	109 words — 6%
2	Dang Pham, Lisa Kaltenegger. "Color classification of Earth-like planets with machine learning", Monthly Notices of the Royal Astronomical Society, 2021 Crossref	61 words — 3%
3	Dang Pham, Lisa Kaltenegger. "Follow the water: Finding water, snow and clouds on terrestrial exoplanets with photometry and machine learning", Monthly Notices of the Royal Astronomical Society: Letters, 2022 Crossref	27 words — 1%
4	thesai.org Internet	22 words — 1%
5	web.archive.org Internet	21 words — 1%
6	www.coursehero.com Internet	18 words — 1%
7	ecommons.cornell.edu Internet	17 words — 1%
8	Siddharth S. Sahu, Vantari Siva, Paresh C. Pradhan, Maheswar Nayak, Kartik Senapati, Pratap K. Sahoo.	15 words — 1%

"Progressive magnetic softening of ferromagnetic layers in multilayer ferromagnet-nonmagnet systems and the role of granularity", Journal of Applied Physics, 2017
Crossref

9 Z Lin, L Kaltenegger. "High-resolution reflection spectra for Proxima b and Trappist-1e models for ELT observations", Monthly Notices of the Royal Astronomical Society, 2019 — 15 words — 1%
Crossref

10 watermark.silverchair.com — 10 words — 1%
Internet

11 www.ddtjournal.com — 10 words — 1%
Internet

12 "Molecular outflows in local (U)LIRGs", Pontificia Universidad Catolica de Chile, 2006 — 8 words — < 1%
Crossref Posted Content

13 Aritra Chakrabarty, Sujan Sengupta. "Generic Models for Disk-resolved and Disk-integrated Phase-dependent Linear Polarization of Light Reflected from Exoplanets", The Astrophysical Journal, 2021 — 8 words — < 1%
Crossref

14 hdl.handle.net — 8 words — < 1%
Internet

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	OFF