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

One of the key factors in identifying life-supporting exoplanets is the study of the exoplanet's atmosphere and the determination of their surface composition. Advanced telescopes like ELTs(Extremely Large Telescopes) and the James Webb Space Telescope(JWST) have been developed with the capability to obtain atmospheric composition measurements of Earth-sized exoplanets. However, the traditional methods for obtaining measurements of atmospheric composition through spectroscopy are resource intensive and time-consuming, which led researchers to explore the possibility of using photometry data to study the planets. This project demonstrates the feasibility of using machine learning algorithms for predicting the surface composition of Earth-like terrestrial exoplanets using photometric flux data. Since, the photometric flux data is anticipated from advanced telescopes in the future, our study will help future telescopes in prioritizing the targets of Exoplanets for spectroscopic analysis. Our approach involves employing a combination of ATMOS, PSG, HELIOS-K and PICASO models to generate the data set and running the SVR, Random Forest and Neural Network Algorithms to achieve the objective. The results show that SVR with linear kernel performs best for small datasets while RFR performs very well for low S/N ratio.

1 Introduction

In recent years, the discovery of exoplanets outside of our own solar system has been a major focus of astronomical research. With advances in telescope technology, astronomers have been able to detect thousands of exoplanets, including many that are similar in size and composition to Earth. However, identifying and characterizing life-supporting exoplanets remains a challenging task. Thus far, the detection of over 5000 exoplanets has been made, with only a handful of these exoplanets located in the habitable zone. This region encircles a star and presents the possibility of liquid water existing on the surface of the planet. Detailed investigations into the atmosphere and surface properties of these planets are necessary to determine whether they are potentially habitable.

One of the methods employed in analyzing an exoplanet's atmosphere involves the use of spectroscopy, which entails observing the spectrum of the exoplanet to look for specific biosignatures. Biosignatures such as methane or oxygen could indicate the existence of life; however, their detection does not always imply the presence of life, as they may have other natural causes.

Another way of detecting life on exoplanets involves identifying surface features that may indicate the presence of living organisms. Previous research has demonstrated that photometric colors of planetary bodies can differentiate between various types of surfaces, such as icy, rocky, or gaseous surfaces. Additionally, models of Earth-like exoplanets tend to fall within a specific color space. Therefore, our work will focus on investigating the surface of exoplanets using the photometric data.

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

2 Exoplanetary surface probing using Machine Learning

Previous studies ((Pham & Kaltenegger, 2022) and (Pham & Kaltenegger, 2021)) have explored the possibility of using broad-band filter photometry combined with machine learning and Markov chain Monte Carlo (MCMC) to detect water in various forms (snow, clouds, and liquid water) on Earth-like exoplanets. Specifically, the XGBoost algorithm was trained for binary classification and to predict the presence of snow, clouds, and water on exoplanetary surfaces. The algorithm demonstrated high balanced accuracy (> 90%) for snow and clouds with a given S/N $\geq$ 20 and up to 70% balanced accuracy for liquid water. Additionally, the study identified the top five optimal filters for identifying these surface features on a terrestrial planet's surface based on the algorithm's feature importance ranking. These filters could be incorporated into telescope designs that search for water on exoplanets.

To test the performance of these optimal filters, Bayesian inferences using MCMC were carried out in an Earth case study with around 100 random realizations of planetary models. Results showed optimal accuracy for snow and clouds at S/N $\geq$ 50 within approximately 5% of the true flux. However, liquid water detection proved more challenging, with most predictions falling within 20% of the true values.

To generate reflection spectra for cold Earth-like exoplanets with modern outgassing rates, a surface temperature of 273K, and a reduced solar spectrum luminosity of 0.875 $L_{\odot}$, an advanced coupled 1D Photochemical Climate model, Exo-Prime2 ((Madden & Kaltenegger, 2020)), in conjunction with a Radiative Transfer component, was employed. The model incorporates key molecules such as $C_2H_6, CH_4, CO, CO_2, H_2CO, H_2O$ necessary for spectroscopic analysis.

The study incorporates surface reflectivity by selecting six major components, namely water, snow, basalt, vegetation (aspen leaf), sand, and clouds, whose albedos are integrated into the model. The MODIS 20 ((King et al., 1997))μm cloud model is used to obtain the albedo of clouds, while the USGS Spectra Library is used for other major surface components. The atmospheric profile is held constant, and nine ideal broad-band filters are implemented to generate true flux values within the wavelength range. Gaussian noise is added to these values to create observed flux values that serve as the data set for the machine learning algorithm.

In a similar study, (Pham & Kaltenegger, 2021) investigates the identification of surface features and biota classification using photometry with Johnson filters. They created a reflection spectra grid for terrestrial earth-like planets with varying surface compositions and cloud coverage, assessing the sensitivity of the results to six different biota samples, including vegetation and UV-resistant biota. The balanced accuracy varied between 50% and 75% for different algorithms and depended on the signal-to-noise ratio.

Baseline Algorithms

In the study by (Pham & Kaltenegger, 2022), the XGBoost algorithm is utilized on simulated photometric data to detect water, snow, and clouds based on characteristic features. The effectiveness of this approach is demonstrated on a set of synthetic exoplanet spectra with various levels of noise and other sources of uncertainty. The accuracy values are compared at different signal-to-noise ratios to evaluate the performance of the algorithm.

Similarly, in (Pham & Kaltenegger, 2021), the efficacy of seven different algorithms is compared for predicting the presence of specific biota on the exoplanet surface using simulated photometric flux with added noise. The classification algorithms include 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 accuracy values are compared for each of these algorithms with varying signal-to-noise ratios.

3 Experiment Review

Github link Our data and algorithms are contained in the link here: ML data and codes.

This report describes an attempt to replicate a previous study by (Pham & Kaltenegger, 2022) that used machine learning (ML) to predict the surface composition of Earth-like exoplanets. The idea is to assess the feasibility of machine learning regression algorithms in predicting the exact surface composition of exoplanets.

To achieve the goal, we will divide the work into two goals:

1st goal: To apply the ML algorithms and the techniques proposed in the paper on the data set already provided in the paper (Pham & Kaltenegger, 2022).

2nd goal: To generate the data set by coupling the the two models Atmos ((Arney et al., 2016)) and PICASO ((Batalha et al., 2019)) instead of using the standard model EXOPRIME-2 ((Pham & Kaltenegger, 2022)).
Finally, combine the generated data set with the avilable data for further training and application of the techniques and algorithms followed in the 1st section to this combined data set.

Data-set description: The dataset used in this study comprises reflection photometric flux data for six different surface combinations, including Sand, Sea water, Vegetation, Cloud, Basalt, and Snow. The flux data is obtained by applying 9 ideal filters of width 0.215 μm to the reflection spectra over a wavelength range spanning from 0.41 μm to 2.35 μm. The spectra is generated by applying the planetary models (EXOPRIME2 (in case of the data provided in the paper (Pham & Kaltenegger, 2022)) and ATMOS coupled to PICASO (in case of the data we planned to generate)).The flux file has two lists in each row, one corresponding to the label i.e the combination of the surface composition and the next list corresponding to the feature i.e. the 9 flux values pertaining to that surface combination. Hence, the dataset covers all possible percentage combinations of these six reflecting surface components with a step size of 5 percent, resulting in a total of 53,130 combinations (refer Table: 1). The 9 flux values obtained upon convoluting the spectra with the 9 filters is shown in Table:2.

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

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 surface combinations and flux values as shown in Table 1 and Table 2 are considered as labels and features respectively for the ML models applied in these data-sets.

Prior work: In the previous report, we made some progress to achieve both the 1st and the 2nd goal listed here.

To achieve the 1st goal, we trained the algorithms using the dataset from the paper. Our approach involved predicting the percentage surface combination of exoplanets using regression algorithms.

Two machine learning algorithms, Support Vector Regression (SVR) and Random forest Regression (RFR), were used to predict the composition from the photometric flux values. The hyperparameters in the case of SVR are the kernel function, while that for RFR is the number of trees considered. The accuracy of the SVR for various kernel functions were compared, and it was found that the linear kernel performed best for the dataset as shown in Figure:9. Hence, we used linear kernel in SVR to the data set. The mean squared error (MSE) obtained without adding noise to the validation dataset for the SVR using the linear kernel for nine filters is 0.00216765. For RFR, the MSE was calculated to be 0.0035305 with the number of trees taken to be 100. The balanced accuracy obtained for classification in the reference paper((Pham & Kaltenegger, 2022)) for 100 signal-to-noise ratio (i.e noiseless signal) for XGBoost was above 95 percent. So, the accuracy we obtained in prediction using noise-less data and applying regression is quite good.

For the 2nd goal of generating data, the Photochemical Climate model, Atmos is to be used along with the Radiative Transfer model, PICASO, to generate simulated photometric flux data for exoplanets. So, in our previous work, The Atmos model ran iteratively until convergence was achieved for the modern earth conditions. Figure:1 and Figure:2 shows atmos finding convergence for temperature and water mixing ratio varying with altitute of the atmosphere. After 6 to 7 iterations, the model seems to have converged completely. Hence the resulting Temperature profiles and Abundances of water and other molecules were then fed into PICASO. Meanwhile, the albedo values for each of the surface components are obtained from the Planetary Spectrum Generator (PSG) as shown in Figure:3 and given as input to PICASO along with the temperature profile and molecular abundance files generated from Atmos. PICASO is equipped with a surface reflectivity module which is used to generate the Reflection Spectra. The albedo values from PSG were in close agreement with the values used in the reference paper((Pham & Kaltenegger, 2022)). However, the opacity file that PICASO used only had a few molecules and did not consider the spectroscopically active molecules for Modern Earth's atmosphere. So, we accounted for that in the later work which is explained in the next section.

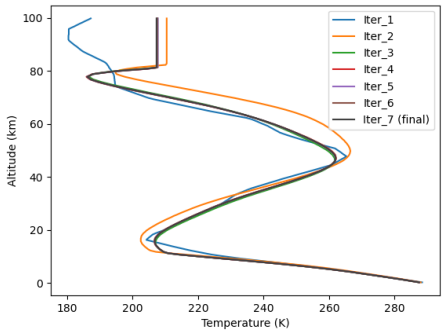


Figure 1: Altitude vs Temperature profile

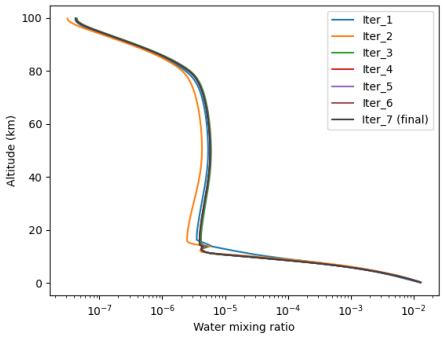


Figure 2: Altitude vs Water Mixing Ratio

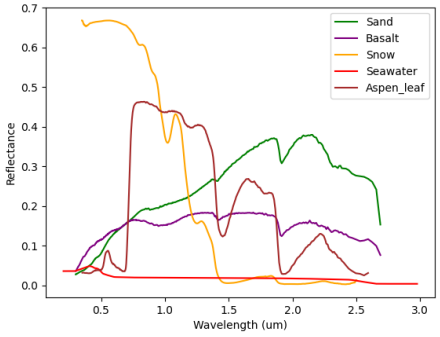


Figure 3: Albedos of components, taken from PSG

4 New Attempts

4.1 Data generation

Modern earth atmosphere was considered for this project which meant an opacity database containing a wide variety of molecules had to be provided to PICASO to obtain the correct transmission spectra. The ready made opacity file databases such as DACE did not have all the molecules that was required. Hence, Helios-K an open source opacity generator had to be used for generating opacities for the spectroscopicaly active molecules (C2H6, CH4, CO, CO2, H2CO, H2O, H2O2, H2S, HNO3, HO2, N2O, N2O5, NO2, O2, O3, OCS, OH, SO2) in the case that is being considered(Modern earth atmosphere). The Helios-K opacity generator is a widely used software tool that is employed in astrophysical simulations to generate opacity tables for a variety of applications, including stellar evolution, supernova explosions, and many other areas of research. Opacity tables play a crucial role in these simulations, as they provide information about the energy transport mechanisms that take place in various astronomical environments. A flowchart of the data-generation is provided in Figure:4

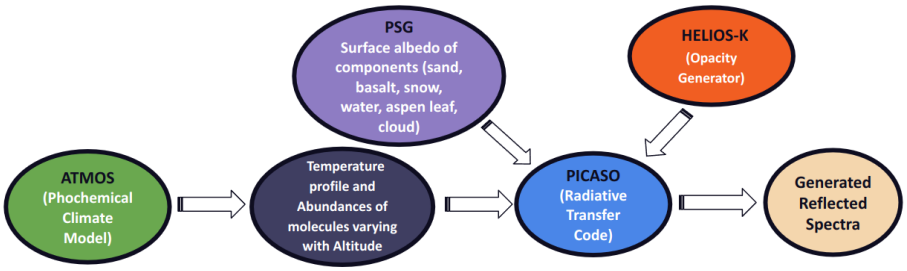


Figure 4: Data Generation Flowchart

The Helios-K opacity generator is designed to calculate and produce opacity tables based on the given thermodynamic conditions, including temperature, density, and chemical composition. One of the key features of the Helios-K opacity generator is its ability to generate opacity tables for a variety of elements and their isotopes. This feature is essential in this project as we need opacities for a wide variety of molecules. The Helios-K opacity generator uses state-of-the-art atomic physics models to compute the opacity of astrophysical materials. These models take into account a range of physical processes, including atomic absorption, electron scattering, and line transitions. The tool also incorporates the latest data from experimental and theoretical studies, ensuring that the computed opacities are as accurate and reliable as possible.

The line lists partition files required as input for HELIOS-K was obtained from HITRAN's line by line database and ExoMol.

4.2 Application of ML algorithms

The nine filter functions were convoluted into the spectra provided in the paper to generate 9 flux values for each combination of the surfaces. The 53130 sets flux data was then divided in the ratioo of 80:20 for the training and validation set respectively. We added 1 noise realization at a particular signal to noise ratio (S/N) to each of the flux combination which then served as our validation set. The following ML algorithms were implemented in the data set and performance(MSE) was analysed. Multi-layer perceptron (MLP) is a supervised neural network model which is capable of learning non-linear models. Implementation of MLP was done using tensorflow library. In this project MLP was trained with the training set and performance was calculated on the validation set for various signal to noise ratios from zero to hundred in steps of ten and compared with other regression algorithms like Support Vector Regressor(SVR), Random Forest Regressor(RFR) and Xgboost. Implementation of SVR and RFR were done using scikit-learn library. Xgboost was run by using Xgboost library. The comparison was done by taking the MSE for each algorithm. Relu activation function was used in the hidden layers and a linear activation function was used at the end. A flowchart for the implementation of the Ml Algorithms is given in Figure: 5

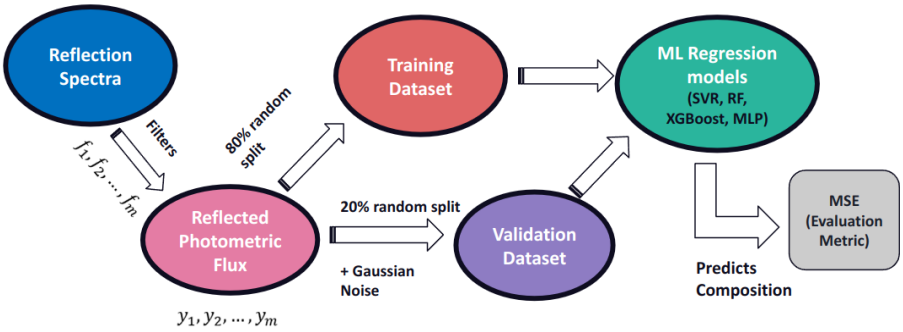
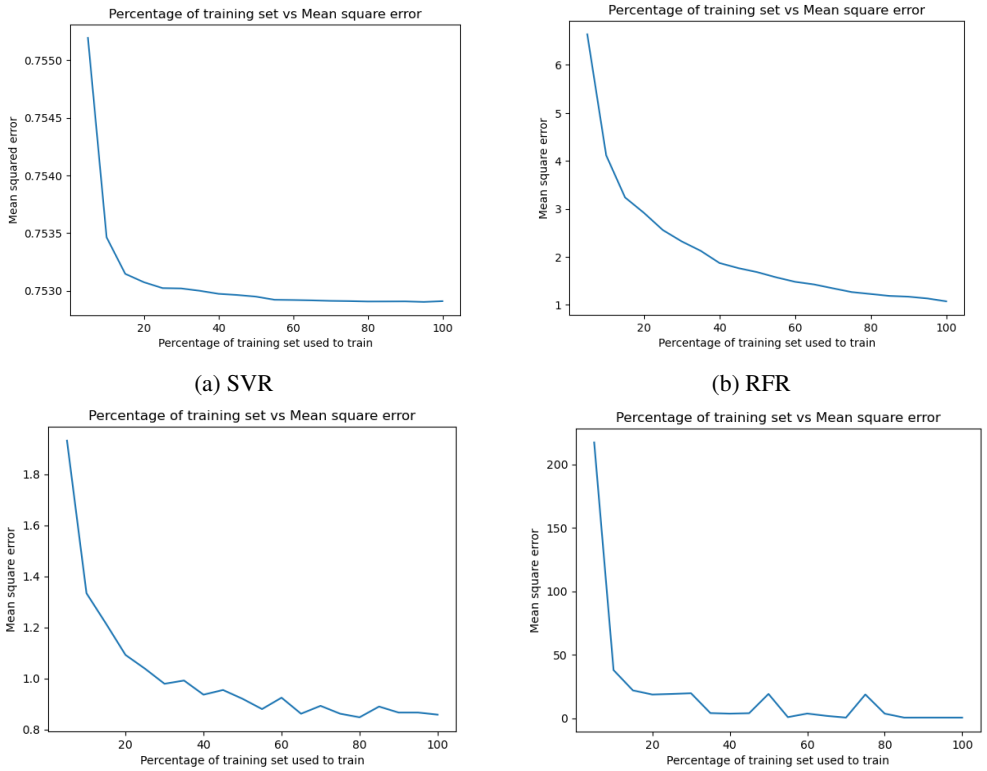


Figure 5: ML Implementation Flowchart

5 Results and Analysis



(a) SVR

(b) RFR

(c) Xgboost

(d) MLP

Figure 6: MSE for different training dataset sizes at S/N ratio=70

The models were trained on noiseless dataset and was tested different S/N. The above figures are shown for S/N = 70. In figure6 we can see how the accuracy of the four algorithms change with training data size. We see that the SVR reaches the saturation point very easily and is performing very well even when it is trained with a very small dataset. It saturates at MSE of around 0.753 and achieves it with just 20% of the training dataset which is around 8500 data points. It performed better than XGBoost whose accuracy never went below 0.8. MLP accuracy was fluctuating a lot with very high MSE for small datasets.

Figure 7: MSE vs Iteration no: MLP

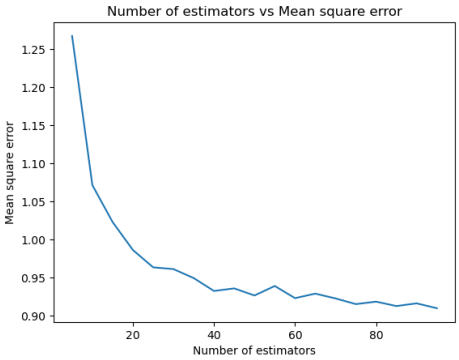


Figure 8: MSE vs number of estimators: RFR

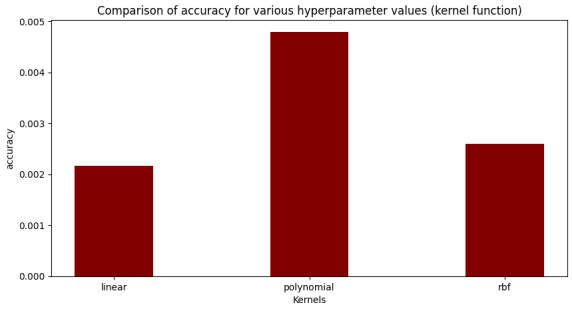


Figure 9: MSE for different kernel functions: SVR

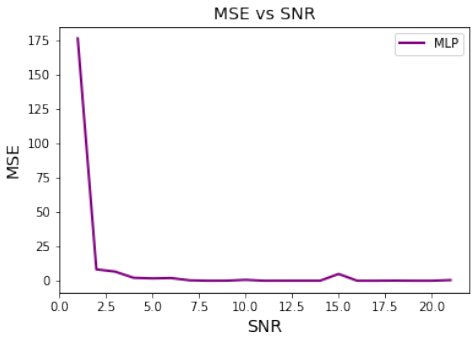


Figure 10: MSE vs SNR(0 to 20) for MLP

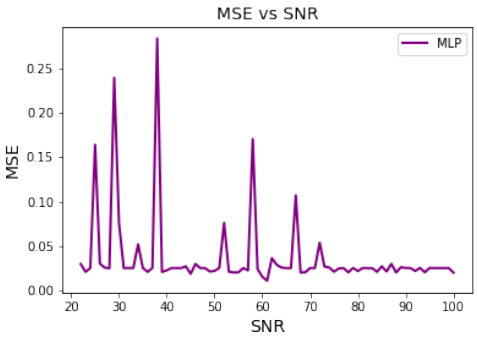


Figure 11: MSE vs SNR(20 to 100) for MLP

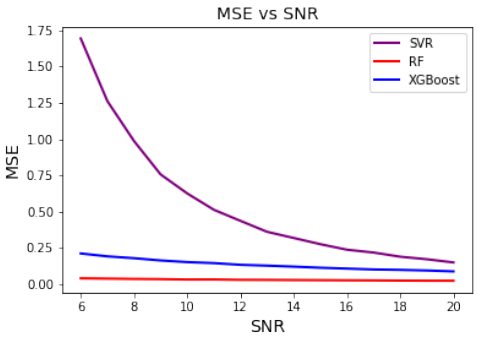


Figure 12: MSE vs SNR(0 to 20) for SVR,RF and XGBoost

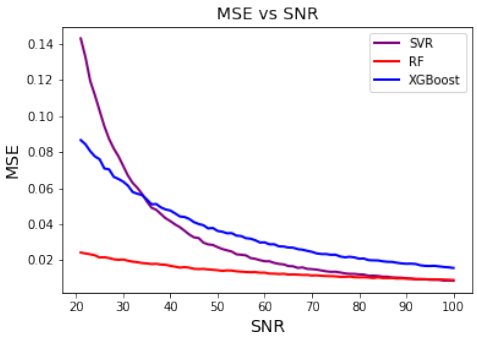


Figure 13: MSE vs SNR(20 to 100) for SVR,RF and XGBoost

In figures 7, 9, 8 in the case of test data having a signal to noise ratio of 70 it can be seen how MSE changes as we change different hyperparameters of each algorithm. In figure 7 we see that that the MSE is fluctuating a lot but stabilizes below 5 after 90. In the case of 8 we see that the MSE almost saturates below 0.95 when number of estimators is above 40. From the bar graph shown in figure 9 we see that SVR performs best when linear kernel is selected with MSE just above 0.02 for S/N=100.

In figures 10, 11, 12, 13 we see that how MSE is varying with S/N ratio. For the case of MLP, in 10 and 11, we see that the MSE goes below 10 for S/N ratio of around 25 then fluctuates between MSE 0.025 and 0.25 between S/N 20 and 100. In figures 12 and 13 we see that Random forest performs best and stays consistent with MSE staying below 0.1 throughout. SVR performs worse than XGBoost for a signal to noise ratio below 35 but overtakes XGBoost for S/N ratio above 35.

6 Conclusion

The results show that machine learning algorithms can predict the surface composition of exoplanets from photometric flux values obtained from their reflection spectra with reasonable accuracy. The accuracy of the SVR using the linear kernel is found to perform very well compared to all the other algorithms which was considered when training dataset is small with reasonable amount of noise(S/N=70). RFR performs best consistently for resonably large datasets with respectable accuracies even for S/N<20. MLP was seen to not perform so well which might be due to dataset being very small. The dataset used in this study covers a wide range of surface combinations and is expected to be useful in the future studies of exoplanets.

7 Further Plan for paper submission

In this section, we outline our future plan for predicting the surface composition of a planet using spectral analysis. Due to technical limitations, we were unable to obtain the albedo for the cloud component. Therefore, we will proceed with the remaining five surface components. The process of generating these molecular opacities takes approximately 2 days to complete. Once complete, we will use these molecular opacities to generate spectra using PICASO. Since PICASO has already been fed with pressure-temperature profiles and abundances, generating data will be swift once the opacities are fed. We will then normalize the generated data. To train the algorithms, we will combine the generated spectra with the available spectra in the paper, following the same steps as we did with the available dataset in this report. These steps involve applying filters to the spectra to get flux, splitting the data into a 20:80 ratio for the training and validation set, and augmenting noise into the validation set. Ultimately, this will allow us to predict the surface composition of the planet. The accuracy of the predictions will be dependent on the quality and availability of the spectral data used for training the algorithms. We aim to improve the accuracy of our method through further refinement and slight tuning of the hyper-parameters. The results of this work will provide valuable insights into the surface makeup of planets and their potential for supporting life.

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