Analysing Exoplanetary Atmospheres using ML

Ayush Singhal **Gaurav Shukla**
School of Physical Sciences
National Institute of Science Education and Research, HBNI, Jatni-752050, India
ayush.singhal@niser.ac.in
gaurav.shukla@niser.ac.in

Abstract

We explored the use of machine learning (ML) techniques to study the atmospheres of exoplanets. The traditional methods for analyzing spectral data from exoplanetary atmospheres rely on manual inspection and interpretation, which can be time-consuming and complex. We proposed using a supervised ML approach to classify and characterize exoplanetary atmospheres based on their spectral features. This demonstrate the effectiveness of the approach by applying it to a dataset of simulated exoplanetary spectra, showing that the ML model can accurately classify the spectra into different atmospheric types and provide estimates of atmospheric properties. The use of ML in the analysis of exoplanetary atmospheres has important implications for the search for habitable exoplanets and the understanding of planetary systems, particularly in the upcoming era of space telescopes such as the James Webb Space Telescope.

1 Introduction

Exoplanets that pass in front of their host star in our line of sight are a small fraction (1%) of the total exoplanet population. Observing them is similar to observing the transit of Venus, but from many light years away. We cannot directly observe the planet itself due to the large distances involved, so we analyze the variations of light coming from the star. This dip in brightness, which is directly proportional to the ratio of the areas of the planet and star, is called a lightcurve.

When stellar light passes through a planet's atmosphere, molecules in the atmosphere can absorb or re-emit different light wavelengths, which leaves a characteristic fingerprint on the light that reaches us. By measuring the change in the dips (transit depth) as a function of wavelength/frequency of light, we can work out which molecules or clouds absorb photons in the atmosphere and understand the planet's chemistry, temperature, cloud coverage, wind speeds, and climate. However, this change in transit depth is only of the order of 0.001% of the light we receive from the star, making this a very challenging observation.

One of the main challenges of studying exoplanetary atmospheres is the complexity of the planetary models required to understand the complex processes happening in their atmospheres, including chemistries, clouds, and dynamics. To overcome the challenges of analyzing spectral data from exoplanetary atmospheres, machine learning (ML) techniques can be used. By using ML algorithms to classify and characterize exoplanetary atmospheres based on their spectral features, we can obtain more reliable and comprehensive results than traditional manual inspection and interpretation methods. ML techniques can also help identify potential candidates for further study and determine which exoplanets may have the necessary conditions for life to exist. With the upcoming launch of space telescopes such as the James Webb Space Telescope, the analysis of exoplanetary atmospheres is expected to enter a new era. However, this increased data volume presents new challenges that require efficient and accurate analysis techniques. In light of these developments, the use of ML techniques to study the atmospheres of exoplanets is a critical area of research that can enhance our understanding

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

of the nature and habitability of exoplanets and contribute to the search for life beyond our solar system.

2 Dataset

We used the dataset used by the authors of the MN18 paper for initial experimentation. The dataset of 100,000 noisy synthetic spectra was generated by using the forward model of Heng & Kitzmann (2017). The spectra were generated in the wavelength range 0.8 - 1.7 μm, and five parameters described each spectrum: temperature (T), volume mixing ratios of water (X_{H2O}), ammonia (X_{NH3}), and hydrogen cyanide (X_{HCN}), and a constant cloud opacity (κ_0). The values of the parameters were chosen randomly from a uniform or log-uniform distribution.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	
0	1.376389	1.374745	1.413392	1.450904	1.466375	1.443057	1.471660	1.432622	1.478647	1.517601	1.533182	1.553194	1.544717	2712.064538	-9.2036
1	1.603904	1.609323	1.628826	1.620339	1.637257	1.658848	1.646935	1.654072	1.682700	1.702930	1.704248	1.688455	1.672496	2392.301318	-0.4577
2	1.478304	1.484133	1.535701	1.549377	1.562368	1.535739	1.546677	1.523513	1.562554	1.594719	1.594487	1.615531	1.614822	1892.056087	-4.7446
3	1.376006	1.374236	1.381826	1.371236	1.372663	1.391107	1.391560	1.388384	1.409598	1.428910	1.426103	1.415693	1.405483	2258.214546	-6.5137
4	1.563088	1.574923	1.570010	1.562674	1.564904	1.566797	1.572265	1.567724	1.556413	1.575183	1.565158	1.564148	1.560290	2752.310725	-10.1175

Figure 1: Generated Dataset containing 13 features and 5 parameters

3 Paper Analysis

3.1 MN18 Paper

The study proposed using supervised machine learning techniques, specifically Random Forest, to classify and characterize exoplanetary atmospheres based on their spectra. The authors explained that using machine learning techniques can significantly improve the efficiency and accuracy of atmospheric characterization, which is crucial for understanding the nature and habitability of exoplanets. The paper begins by briefly explaining exoplanetary atmosphere characterization techniques and the challenges of interpreting spectral data. The authors note that traditional methods for analyzing spectral data rely on manual inspection and interpretation, which can be time-consuming and subjective. Additionally, the complexity of the data and the noise present in observations can make it difficult to identify patterns and trends.

The authors describe their methodology for using regression trees and bootstrapping to analyze a dataset of synthetic spectra. They explain that they randomly draw from the training set of 80,000 synthetic spectra to train each regression tree, and that each drawn spectrum is placed back into the training set, allowing for it to be drawn more than once. They note that a single regression tree produces predictions with large uncertainties, but that these uncertainties can be mitigated by combining the responses of multiple trees in a random forest. They performed tests to ensure the convergence of the predictions using 1000 regression trees, which allowed them to compute the posterior distributions of the parameters. Overall, this methodology allows for the computation of the posterior distributions of the parameters for the given data points.

They trained their model on 80,000 synthetic spectra and used it to analyze 20,000 more synthetic spectra. They found that the outcomes of the retrievals converged when the number of trees used exceeded 100. They also tested the retrieval outcomes with different levels of assumed noise floors, which represent the uncertainty in the transit depths of the data points in the synthetic WFC3 spectra. They found that the variance associated with the true versus predicted values of the parameters decreased when the assumed noise floor was lower. Overall, these tests demonstrate the robustness of the authors' implementation of the random forest method for analyzing the synthetic spectra.

4 Experiments

4.1 Experiment 1

Dataset was provided for the MN18 paper and we used that to train the Random Forest Model using the 80% split of the training data and using other 20% for testing the model and then used that model

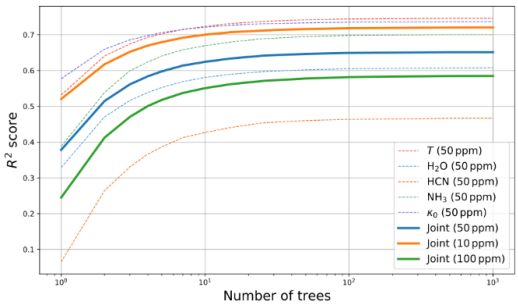

Figure 2: This shows the coefficient of determination (also known as R^2) for each of the 5 parameters and for the joint prediction, as a function of the number of regression trees used in the random forest which helps to evaluate the accuracy of the model predictions as a function of the number of regression trees used in the random forest, as well as the assumed noise floor of the data.(Ref. = MN18)

to predict the composition of the planet WASP 12-b whose binned data was also provided. Then we used the binned data for a newer planet HD209548b and used the earlier trained model for the prediction of the composition of the planet's atmosphere.

```
Prediction for $T (K)$: 892 [+421 -145]
Prediction for H$_2$O: -2.34 [+1.6 -3.12]
Prediction for HCN: -7.52 [+3.97 -3.6]
Prediction for NH$_3$: -9.3 [+4.39 -3.1]
Prediction for $\kappa_0$: -2.35 [+1.4 -1.32]
```

Figure 3: Using the model trained on the dataset from the MN18 paper to predict the atmospheric composition of the planet WASP12-b
(Predicted values alongwith $[+1\sigma\,,\,-1\sigma]$)

4.2 Experiment 2

Pycaret is an open-source machine learning library for Python that enables users to perform end-to-end machine learning experiments quickly and efficiently. The package includes many pre-processing and modeling functions, allowing users to build and tune models with just a few lines of code.

The Pycaret package was used to find the best algorithm for a regression problem and it was determined that the most suitable algorithms for the given dataset are Extra Trees Regressor and Random Forest Regressor, this suggests that the data has complex relationships and the chosen algorithms are capable of handling such complexity. Extra Trees Regressor and Random Forest Regressor are both ensemble algorithms that use a collection of decision trees to make predictions. They have a reputation for being robust and accurate models that perform well on a variety of datasets. The choice between Extra Trees Regressor and Random Forest Regressor may depend on factors such as the size of the dataset, the number of features, and the desired level of interpretability. Extra Trees Regressor is known for its fast training time and can work well on small datasets, whereas

```
Prediction for $T (K)$: 1.32e+03 [+967 -492]
Prediction for H$_2$O: -7.12 [+4.56 -4.33]
Prediction for HCN: -7.12 [+3.58 -3.75]
Prediction for NH$_3$: -11.7 [+7.03 -1.34]
Prediction for $\kappa_0$: -1.81 [+2.27 -1.63]
```

Figure 4: Using the model trained on the dataset from the MN18 paper to predict the atmospheric composition of the planet HD209548b
(Predicted values alongwith $[+1\sigma\,,\,-1\sigma]$)

Random Forest Regressor is often more accurate but can be computationally expensive and slower to train. Therefore, the specific characteristics of the dataset and project goals should be taken into consideration when deciding which algorithm to use for the regression problem.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
et	Extra Trees Regressor	238.6207	124924.6034	353.4246	0.7390	0.2548	0.1829	2.3830
rf	Random Forest Regressor	241.1384	127595.1316	357.1680	0.7334	0.2570	0.1837	6.6150
lightgbm	Light Gradient Boosting Machine	262.3880	135255.8088	367.7395	0.7174	0.2618	0.1961	0.0970
knn	K Neighbors Regressor	244.4891	139910.3516	374.0161	0.7077	0.2676	0.1826	0.0990
gbr	Gradient Boosting Regressor	310.1038	167458.8657	409.1884	0.6501	0.2888	0.2294	2.4940
ada	AdaBoost Regressor	410.2363	244564.0773	494.5090	0.4891	0.3681	0.3405	0.4080
lr	Linear Regression	407.4809	259394.4266	509.2784	0.4581	0.3607	0.3106	0.5150
br	Bayesian Ridge	407.4955	259394.4003	509.2784	0.4581	0.3607	0.3107	0.0290
dt	Decision Tree Regressor	330.1323	259622.7330	509.4812	0.4575	0.3544	0.2452	0.1360
ridge	Ridge Regression	412.5893	262361.7094	512.1870	0.4519	0.3643	0.3163	0.0180
huber	Huber Regressor	398.6638	269333.4790	518.9351	0.4373	0.3566	0.2782	0.4110
lasso	Lasso Regression	438.0545	286980.4094	535.6846	0.4004	0.3831	0.3409	0.2650
par	Passive Aggressive Regressor	411.6459	289429.3807	537.6923	0.3954	0.3714	0.2830	0.2100
omp	Orthogonal Matching Pursuit	484.5532	351491.4421	592.8413	0.2657	0.4267	0.3843	0.0180
lar	Least Angle Regression	498.9884	393108.0198	626.9096	0.1788	0.4519	0.3633	0.0200
llar	Lasso Least Angle Regression	547.1923	401893.8004	633.9365	0.1604	0.4482	0.4440	0.0170
en	Elastic Net	582.5949	451872.3406	672.2014	0.0560	0.4697	0.4738	0.0190
dummy	Dummy Regressor	599.6262	478804.8344	691.9437	-0.0002	0.4808	0.4881	0.0170

Figure 5: Using the PyCaret to find the best suitable model for the problem

Github link for our code and data : https://github.com/LuminAYUSH/ML_Project_group9

5 Future plans

We want to use the Dataset from the Ariel ML Data Challenge which is generated with Alfnoor, which combines the open source TauREx 3 atmospheric modelling suite with the official Ariel instrument simulator ArielRad to produce large-scale simulations of atmospheres.

We also want to use extra tress regressor because its faster,less compute heavy and best suites the type of dataset we are using.And also tune the hyperparameters to get the best accuracy from both the random forest regressor and extra tree regressor.

Explore the possibility to apply neural networks if time permits.

References

[1] Márquez-Neila, Pablo et al. “Supervised machine learning for analysing spectra of exoplanetary atmospheres.” Nature Astronomy 2 (2018): 719-724.

[2] Nixon, Matthew C. and Nikku Madhusudhan. “Assessment of supervised machine learning for atmospheric retrieval of exoplanets.” Monthly Notices of the Royal Astronomical Society 496 (2020): 269-281.

[3]Munsaket, Patcharawee; Awiphan, Supachai; Chainakun, Poemwai; Kerins, Eamonn "Retrieving exoplanet atmospheric parameters using random forest regression. " Journal of Physics: Conference Series, Volume 2145, Siam Physics Congress 2021 (SPC 2021) 24-25 May 2021 Thailand

[4] pycaret.org. PyCaret, April 2020. URL https://pycaret.org/about. PyCaret version 1.0.0.