REPORT ON HARNESSING ML FOR ATMOSPHERIC RETRIEVAL OF EXOPLANETS

Swastik Dewan
Tasneem Basra Khan
National Institute of Scientific Education and Research
swastik.dewan@niser.ac.in
tasneembasra.khan@niser.ac.in

Abstract

The Report consists of the mid-way updates of the project on Harnessing ML for Atmospheric Retrieval of Exoplanets. Atmospheric Retrieval is a method to go deep inside the components present in the atmosphere of the exoplanets, to perform the atmospheric retrieval various traditional methods are used but lead to computational complexity so using machine Learning we can increase performance and generate transmission spectra in a much computationally efficient way. HELA which uses random forest is used here to train and test the data set and retrieval is performed.

1 Targets Achieved:

- Completing the literature review about the Atmospheric Retrieval and details about the Radiative Transfer equation.
- Training and testing the already existing Models (HELA (uses Random forest), POSEIDON(Non-ML-nested sampling method)). Both are Python-based models.

2 Literature Review

2.1 Introduction

Exoplanets are planets that orbit stars other than the Sun outside the solar system.(1) The detection of exoplanets and subsequently studying their atmospheric properties such as the chemical compositions, temperature profiles, clouds/hazes, and energy circulation make up a fascinating area of astronomy, in part because the search for worlds orbiting stars other than our Sun provides a unique opportunity to understand the formation of our solar system's planets and the possible end of our own (Madhusudan,2018).

An exoplanet's spectrum offers a glimpse of its atmosphere. The several interrelated physicochemical processes and characteristics of the atmosphere that are disclosed by their impact on the radiation that emerges from the atmosphere and reaches the observer are encoded in a spectrum. In turn, the retrieved attributes can shed light on the many physical and chemical processes that affect the atmosphere and their development history.

The process of characterizing planetary atmospheres involves determining which model parameters best fit the observed exoplanet spectra. Atmospheric retrieval is primarily concerned with connecting exoplanet spectra to the parameters of intricate forward models of atmospheric physicochemical processes.

Bayesian inference methods such as MCMC, nested sampling, and optimal estimation algorithms have been widely used for exoplanet retrieval. The problem with traditional sampling methods is the huge computational time associated with processing observational data, thus machine learning can speed up the process without compromising the accuracy. many such Machine learning models are available like HELA (random forest), INARA, ExoGan (Neural Networks), etc. Once trained the model gives a much faster result than the traditional methods.(Vasist,2023)

2.2 Radiative Transfer Equation

$$t_{\lambda,i} = \mathrm{EXP}\left(-\sum_{j=1}^{N_{\mathrm{lay}}} \alpha_{\lambda,j}\mathcal{P}_{i,j}\Delta h_j\right) = \mathrm{EXP}\left(-\sum_{j=1}^{N_{\mathrm{lay}}} \Delta\tau_{\lambda,j}\mathcal{P}_{i,j}\right),$$

Figure 1: Radiative Transfer Equation : Transit depth

$$\left(\frac{R_{\mathrm{p},\lambda}}{R_{\mathrm{s}}}\right)^2 = \frac{1}{R_{\mathrm{s}}^2}\left(R_{\mathrm{p}}^2 + 2\sum_{i=1}^{N_{\mathrm{r}}}[1 - t_{\lambda,i}]b_i\Delta b_i\right),$$

Figure 2: Radiative Transfer Equation to be used.

This equation shows how a linear ray passes across the atmosphere of the exoplanet. log of Transmission is the summation of path distribution over the atmospheric layers. From the transmission, we can evaluate the radius of the exoplanet.

3 Methodology

3.1 HELA

We have used the HELA model to run the training using 80,000 transmission spectra and 20,000 data for testing the data. The plot R^2 Score which checks the accuracy between the predicted and real value is given below, The $R^2 Score = 1$ is considered to be the most accurate.

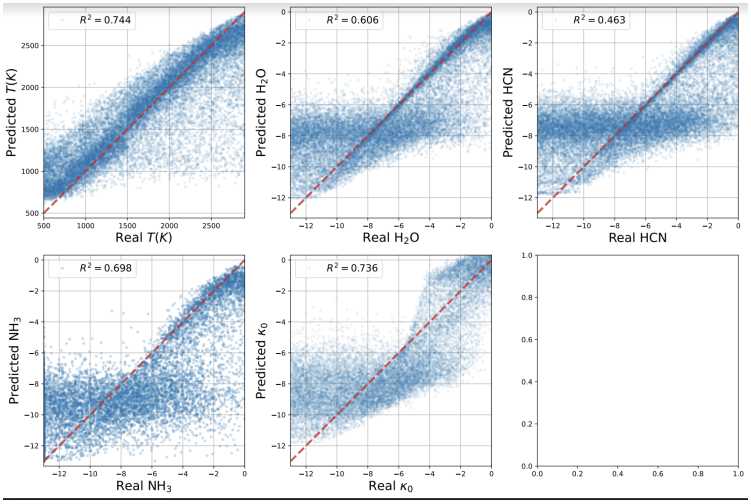

Figure 3: R^2 score after performing the testing.

We also ran the retrieval, in which 1000 regression trees were formed, each regression tree gives a set value range of the parameter. Here 5 parameters were used. The Parameters used include temperature, H2O, HCN, NH3, and specific gravity. The parameter space radius of the exoplanet can be found by using the radiative transfer equation. In Hela, the closest transit generated is considered for calculation and the respective parameter value ranges are the output of the final transmission spectrum.

The output is generated in the form of corner plots as shown below:

Each corner plot gives the range value of the parameters used in this case the given 5.

3.2 Comparison between Different ML models

Our project involves comparative approach to give a comparative analysis of Atmospheric retrieval done using different ML models. There exists no standard data to compare which is the most accurate and efficient method of Atmospheric retrieval.One model works fine for fulfilling certain purpose and other model suits the other purpose. We would like to give users the option of choosing the algorithm based on their needs. Moreover Hela Data set only contains data for few parameters across small wavelength range and the data is specific to hot Jupiter (gaseous exoplanets), we will be adding more in the data set which will be obtained from NISER and also synthetically generated from NASA psg. ‘

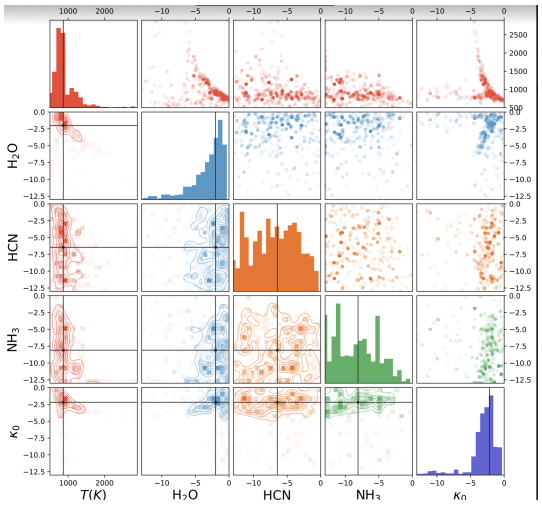

Figure 4: Corner plot of retrieval run in HELA.

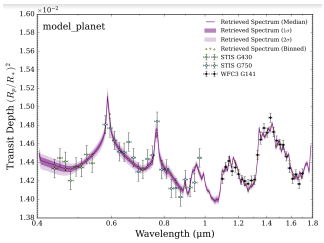

Figure 5: Snapshot of the forward model in POSEIDON which is a Non ML based Algorithm

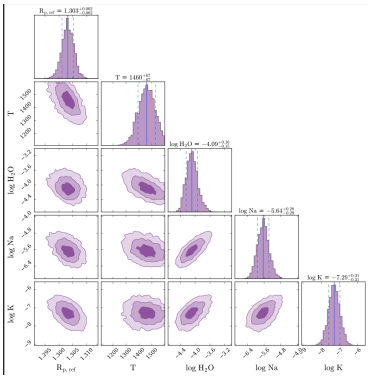

Figure 6: Corner plot of the POSEIDON run retrieval.

Machine Learning Technique	R^{2} Score
1. Random Forest	0.6394
2.XGBRegressor	0.5925
3. SVR	0.5668
4. Neural Networks	0.1456

Figure 7: Comparison of different ML models using HELA data set. USing the data these ML models were run in a very raw form just to check the R^2 score.

4 Further plans

- Data Set would be taken from HELA DATA SET (80,000 2. WFC3 transmission spectra for training and 20,000 datasets for testing) + Data Synthesized in NISER + Data Self - Generated from NASA Psg..
- Creating a Comparative Model of various ML techniques which gives user option to compare the accuracy of respective Atmospheric Retrievals.
- Implementing a Ensemble Learning approach to enhance the accuracy if possible.

5 References:

1. Robinson, T. D. (2017, February 20). A Theory of Exoplanet Transits with Light Scattering. The Astrophysical Journal, 836(2), 236.
2. Márquez-Neila, P., Fisher, C., Sznitman, R., Heng, K. (2018, June 25).Supervised machine learning for analyzing spectra of exoplanetary atmospheres.Nature Astronomy, 2(9), 719–724.
3. MacDonald, R. J. (2023, January 13). POSEIDON: A Multidimensional Atmospheric Retrieval Code for Exoplanet Spectra. Journal of Open Source Software, 8(81), 4873.
4. Hayes, J. J. C., Kerins, E., Awiphan, S., McDonald, I., Morgan, J. S., Chuanraksasat, P., Komonjinda, S., Sanguansak, N., Kittara, P. (2020, April 14). Optimizing exoplanet atmosphere retrieval using unsupervised machine-learning classification. Monthly Notices of the Royal Astronomical Society, 494(3),4492–4508. https://doi.org/10.1093/mnras/staa978
5. All the codes used for training and testing (in HELA) can be found here.

Tasneem Basra Khan paper check

By Tasneem Basra Khan

WORD COUNT	1063	TIME SUBMITTED	13-MAR-2024 02:31PM
		PAPER ID	107532733

REPORT ON HARNESSING ML FOR ATMOSPHERIC RETRIEVAL OF EXOPLANETS

Swastik Dewan
Tasneem Basra Khan
National Institute of Scientific Education and Research
swastik.dewan@niser.ac.in
tasneembasra.khan@niser.ac.in

Abstract

The Report consists of the mid-way updates of the project on Harnessing ML for Atmospheric Retrieval of Exoplanets. Atmospheric Retrieval is a method to go deep inside the components present in the atmosphere of the exoplanets, to perform the atmospheric retrieval various traditional methods are used but lead to computational complexity so using machine Learning we can increase performance and generate transmission spectra in a much computationally efficient way. HELA which uses random forest is used here to train and test the data set and retrieval is performed.

1 Targets Achieved:

- Completing the literature review about the Atmospheric Retrieval and details about the Radiative Transfer equation.
- Training and testing the already existing Models (HELA (uses Random forest), POSEIDON(Non-ML-nested sampling method)). Both are Python-based models.

2 Literature Review

2.1 Introduction

Exoplanets are planets that orbit stars other than the Sun outside the solar system.(1) The detection of exoplanets and subsequently studying their atmospheric properties such as the chemical compositions, temperature profiles, clouds/hazes, and energy circulation make up a fascinating area of astronomy, in part because the search for worlds orbiting stars other than our Sun provides a unique opportunity to understand the formation of our solar system's planets and the possible end of our own (Madhusudan,2018).

An exoplanet's spectrum offers a glimpse of its atmosphere. The several interrelated physicochemical processes and characteristics of the atmosphere that are disclosed by their impact on the radiation that emerges from the atmosphere and reaches the observer are encoded in a spectrum. In turn, the retrieved attributes can shed light on the many physical and chemical processes that affect the atmosphere and their development history.

The process of characterizing planetary atmospheres involves determining which model parameters best fit the observed exoplanet spectra. Atmospheric retrieval is primarily concerned with connecting exoplanet spectra to the parameters of intricate forward models of atmospheric physicochemical processes.

Bayesian inference methods such as MCMC, nested sampling, and optimal estimation algorithms have been widely used for exoplanet retrieval. The problem with traditional sampling methods is the huge computational time associated with processing observational data, thus machine learning can speed up the process without compromising the accuracy. many such Machine learning models are available like HELA (random forest), INARA, ExoGan (Neural Networks), etc. Once trained the model gives a much faster result than the traditional methods.(Vasist,2023)

2.2 Radiative Transfer Equation

$$t_{\lambda,i} = \mathrm{EXP}\left(-\sum_{j=1}^{N_{\mathrm{lay}}} \alpha_{\lambda,j} \mathcal{P}_{i,j} \Delta h_j\right) = \mathrm{EXP}\left(-\sum_{j=1}^{N_{\mathrm{lay}}} \Delta\tau_{\lambda,j} \mathcal{P}_{i,j}\right).$$

Figure 1: Radiative Transfer Equation : Transit depth

$$\left(\frac{R_{p,\lambda}}{R_s}\right)^2 = \frac{1}{R_s^2}\left(R_p^2 + 2\sum_{i=1}^{N_c}[1 - t_{\lambda,i}]b_i \Delta b_i\right).$$

Figure 2: Radiative Transfer Equation to be used.

This equation shows how a linear ray passes across the atmosphere of the exoplanet. log of Transmission is the summation of path distribution over the atmospheric layers. From the transmission, we can evaluate the radius of the exoplanet.

3 Methodology

3.1 HELA

We have used the HELA model to run the training using 80,000 transmission spectra and 20,000 data for testing the data. The plot R^2 Score which checks the accuracy between the predicted and real value is given below, The $R^2 Score = 1$ is considered to be the most accurate.

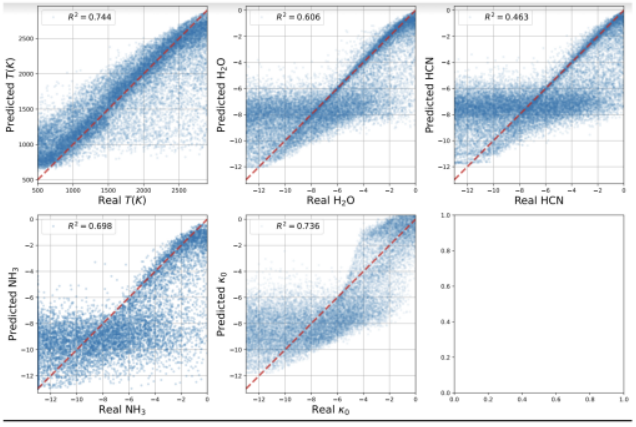

Figure 3: R^2 score after performing the testing.

We also ran the retrieval, in which 1000 regression trees were formed, each regression tree gives a set value range of the parameter. Here 5 parameters were used. The Parameters used include temperature, H2O, HCN, NH3, and specific gravity. The parameter space radius of the exoplanet can be found by using the radiative transfer equation. In Hela, the closest transit generated is considered for calculation and the respective parameter value ranges are the output of the final transmission spectrum.

The output is generated in the form of corner plots as shown below:

Each corner plot gives the range value of the parameters used in this case the given 5.

3.2 Comparison between Different ML models

Our project involves comparative approach to give a comparative analysis of Atmospheric retrieval done using different ML models. There exists no standard data to compare which is the most accurate and efficient method of Atmospheric retrieval.One model works fine for fulfilling certain purpose and other model suits the other purpose. We would like to give users the option of choosing the algorithm based on their needs. Moreover Hela Data set only contains data for few parameters across small wavelength range and the data is specific to hot Jupiter (gaseous exoplanets), we will be adding more in the data set which will be obtained from NISER and also synthetically generated from NASA psg. ‘

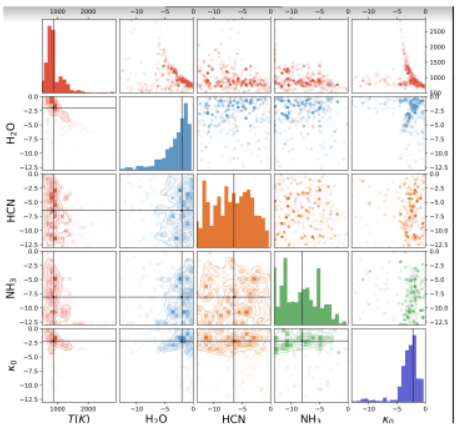

Figure 4: Corner plot of retrieval run in HELA.

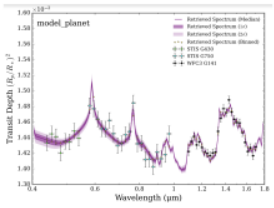

Figure 5: Snapshot of the forward model in POSEIDON which is a Non ML based Algorithm

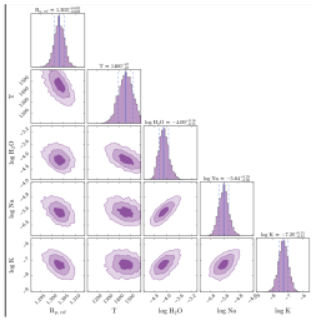

Figure 6: Corner plot of the POSEIDON run retrieval.

Machine Learning Technique	R^{2} Score
1. Random Forest	0.6394
2.XGBRegressor	0.5925
3. SVR	0.5668
4. Neural Networks	0.1456

Figure 7: Comparison of different ML models using HELA data set. USing the data these ML models were run in a very raw form just to check the R^2 score.

4 Further plans

- Data Set would be taken from HELA DATA SET (80,000 2. WFC3 transmission spectra for training and 20,000 datasets for testing) + Data Synthesized in NISER + Data Self - Generated from NASA Psg..
- Creating a Comparative Model of various ML techniques which gives user option to compare the accuracy of respective Atmospheric Retrievals.
- Implementing a Ensemble Learning approach to enhance the accuracy if possible.

5 References:

1. Robinson, T. D. (2017, February 20). A Theory of Exoplanet Transits with Light Scattering. The Astrophysical Journal, 836(2), 236.
2. Márquez-Neila, P., Fisher, C., Sznitman, R., Heng, K. (2018, June 25).Supervised machine learning for analyzing spectra of exoplanetary atmospheres.Nature Astronomy, 2(9), 719–724.
3. MacDonald, R. J. (2023, January 13). POSEIDON: A Multidimensional Atmospheric Retrieval Code for Exoplanet Spectra. Journal of Open Source Software, 8(81), 4873.
4. Hayes, J. J. C., Kerins, E., Awiphan, S., McDonald, I., Morgan, J. S., Chuanraksasat, P., Komonjinda, S., Sanguansak, N., Kittara, P. (2020, April 14). Optimizing exoplanet atmosphere retrieval using unsupervised machine-learning classification. Monthly Notices of the Royal Astronomical Society, 494(3),4492–4508. https://doi.org/10.1093/mnras/staa978
5. All the codes used for training and testing (in HELA) can be found here.

Tasneem Basra Khan paper check

ORIGINALITY REPORT

9%
SIMILARITY INDEX

PRIMARY SOURCES

1	"Handbook of Exoplanets", Springer Science and Business Media LLC, 2018 Crossref	53 words — 6%
2	astrobiology.com Internet	10 words — 1%
3	Yuxiang Yan, Xianwen Yu, Fengyang Long, Yanfeng Dong. "A Multi-Criteria Evaluation of the Urban Ecological Environment in Shanghai Based on Remote Sensing", ISPRS International Journal of Geo-Information, 2021 Crossref	9 words — 1%
4	Schunck, M., M. Hegmann, and E. Sedlmayr. "The influence of stochastic density fluctuations on the infrared emissions of interstellar dark clouds", Monthly Notices of the Royal Astronomical Society, 2007. Crossref	8 words — 1%

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