MLARE : ML Driven Atmospheric Retrieval of Exoplanets

Swastik Dewan
School of Physical Sciences
NISER Bhubaneswar
Jatani-752050
swastik.dewan@niser.ac.in

Tasneem Basra Khan
School of Physical Sciences
NISER Bhubaneswar
Jatani - 752050
tasneembasra.khan@niser.ac.in

Abstract

The paper illustrates the use of Machine Learning in Atmospheric Retrieval of Exoplanets focusing mainly on Hot Jupiter WFC3 transmission spectra. We introduce here a Stacking Ensemble-learning model having base models as Random forest , Gradient Boosting and K-Nearest Neighbours, and meta model as Ridge regression, bench marked with other pre-existing machine learning models. R^2 score was considered the accuracy parameter. For our Ensemble learning approach R^2 score was found out to be 0.752 with added 50ppm and 100ppm noise and bench marked with HELA Atmospheric Retrieval in which the R^2 was 0.676, 0.651 and 0.586 for noise floors of 10, 50 and 100 ppm. Besides that the Data curation was also done using ExoCTK Generic Grid (generation of 112640 data , 5000 transit depth values with 8 labels), The main purpose of data generation was checking if the model prediction was accurate or not and also to increase the parameter space. In the Future works we will be working with the synthetic data generated using ExoCTK to improve on the Retrieval model in terms of accuracy and computational cost.**All the codes used for our project can be found here**.

1 Introduction

1.1 What are Exoplanets

Exoplanets are extra solar planets that orbit stars other than the Sun outside the solar system[5].Studying exoplanets is one of the flagship mission of the current decade. Studying earth-like planets around sun-like stars will not only give us the information about whether life exists outside Earth or not, but also about the stars around which the exoplanets revolve. As the stars would be in different phases of their life cycle, so we can study the evolution of stars. We have more than 5000 exoplanets discovered so far. The detection of exoplanets and subsequently studying their atmospheric properties such as the chemical compositions, temperature-pressure 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 [6].

An exoplanet's spectrum offers a glimpse of its atmosphere. The several interrelated physio-chemical 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. Based on the composition(Mixing Ratios, equilibrium or non equilibrium chemistry), the exoplanets can be classified as Hot Jupiters , Rocky/terrestrials , sub-Neptunes etc. In this paper we mainly consider the data of hot Jupiters (like WASP39b), these are closer to the star and have masses similar or greater to the Jupiter. The studies on these exoplanets are also based on which kind of resolution of telescopes we are using. A higher resolution telescope gives a better wavelength scale and we can get a wider information about the atmosphere. We will mainly consider JWST data and WFC3 (Wide field Camera) resolutions here.

1.2 Atmospheric Retrieval: a review

Atmospheric Retrieval is the technique which helps us to know about the constituents of chemicals or other properties hidden in the intricacies of exo-planetary atmospheres. The atmosphere may contain different sets of molecules and the absorption of light due to those molecules defines the opacities. The reflection of light due to the molecules is called the albedo. Moreover its the transmission, absorption or reflection of light from the exoplanet that reaches the telescope, is what used to predict the molecules and properties. The relationship between different properties is given by the Radiative transfer equation. The atmosphere can be divided into different layers to get the Temperature-Pressure profile across different layers and different phenomenon like collision-induced absorption, pressure broadening come into play. Based on the dynamics of atmosphere we can have equilibrium chemistry or have non equilibrium chemistry which is mainly due to convection or wind in the atmosphere and that is why we will be characterising exoplanets(properties like metallicity, C/O ratio, cloud, haze and chemical abundances like Water, Ammonia etc.) to study the atmosphere of the exoplanet to know about its components and the phenomenon that drives their abundances. Our main focus would be finding the chemical abundance, effective temperature i.e the average temperature of the exoplanet, metallicity i.e how much metal-like elements are there, C/O ratio i.e how much of carbon oxygen is present, constant cloud opacity which will tell us more about the opacity of transmission.

1.3 Some Previous works

Various traditional methods have been used so far including Bayesian inference methods such as MCMC, nested sampling(PyMultinest , Ultranest etc). Already made packages are available like Poseidon [7], Picaso etc which do not use machine learning. Here we would like to introduce Machine learning for increasing the accuracy and bringing up efficient computational power. Many Atmospheric retrieval methods using Machine learning do exist in the community,but are not widely used. The first machine learning technique which came used Random forest method and the Model was called HELA [4]. The prediction is specific to the type of exoplanet, it can be a Hot Jupiter or terrestrial exoplanet and it can be either transmission spectra or emission or reflection spectra. Our main focus would be Prediction of hot Jupiters on the transmission spectra data. There are few Machine Learning models using that category namely HELA [11] , ExoGAN [8] , Plan net [9], Fisher , Madhusudhan , ExoCNN [10], VI retrieval. HELA as said earlier uses Random Forest, ExoGAN uses Generative Adversial Network , Plan net uses Bayesian Neural network, Fisher Madhusudan have worked on Random forest , ExoCNN is a Convulational neural network technique and we also variational inference as well. We will working on JWST/WFC3 Transmission spectra for hot Jupiters and our main focus would be benchmarking the previous technique especially the technique which uses random forest like HELA and Madhusudhan. We have generated the corner plots after running the already existing ones namely of POSEIDON as shown in Figure2 and HELA as shown in the Figure 3, we expect the comparable R^2 score to HELA as shown in Figure 1.

2 Methodology

2.1 Dataset Used for Training

2.1.1 HELA dataset: WFC3 Transmission Spectra

In this paper we are first using data set already available in the community. Many other papers have also used the same data set(HELA data set). The training set consists of 80,000 synthetic WFC3 transmission spectra, each described by 5 parameters: Temperature (T), volume mixing ratios (relative abundances by number) of water (X_{H_2O}), ammonia (X_{NH_3}) and hydrogen cyanide (X_{HCN}), and a constant cloud opacity (κ_0). The training set resides in a 13-dimensional space, where each dimension corresponds to a wavelength bin. Along the axis of each dimension is a continuous range of values of the transit radii. The testing dataset has 20,000 synthetic spectra, arranged in similar fashion. HELA dataset is a synthetically generated WFC3 dataset. The data set can be found in the HELA github repository [11]. Before training we have used standardscaler(), besides that robustscaler()

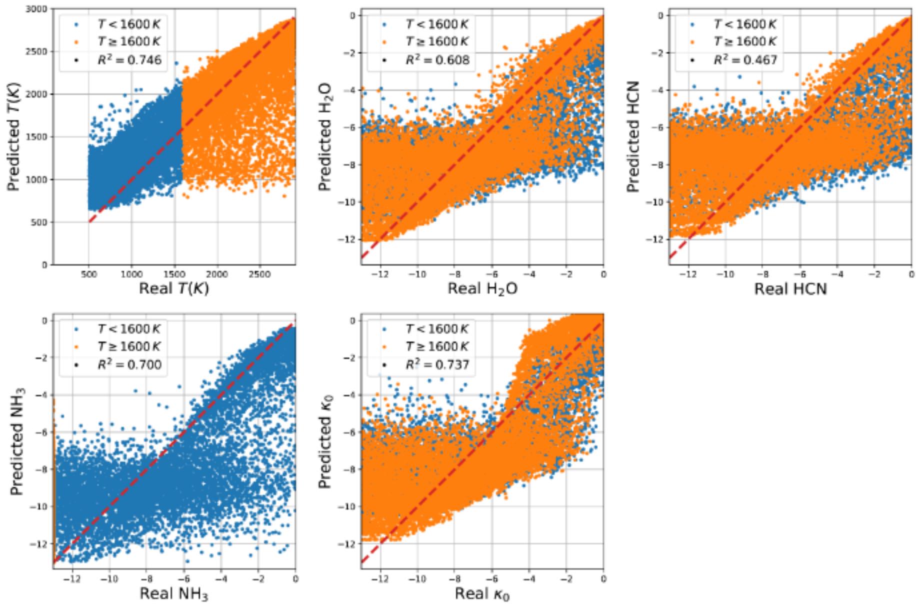

Figure 1: R^2 prediction of Individual parameter (HELA) from the HELA Paper [11]

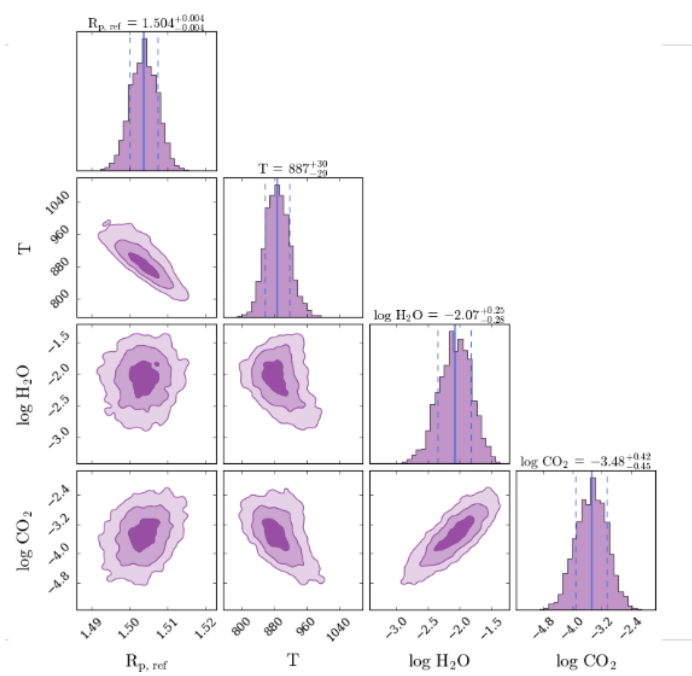

Figure 2: Corner plot for POSEIDON Retrieval from the POSEIDON Paper [7]

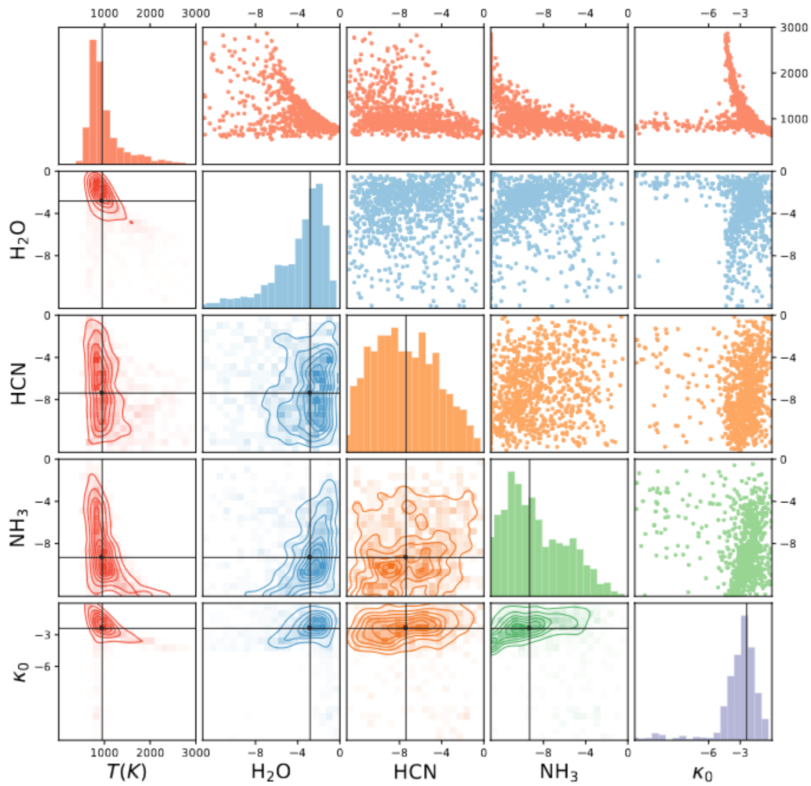

Figure 3: Corner plot for HELA Retrieval, 5 parameters are being retrieved from the HELA paper [11].

and MinMaxscaler() was also used to scale the data but gave same results. Noise was also added as 50ppm training and 100ppm testing to avoid over fitting. In the HELA paper noise was assumed to be a Gaussian uncertainty on the transit depths with full widths at half-maximum of 10, 50 and 100 ppm which represent ideal, typical and easily attainable conditions. As a further check they assumed noise floors of 10, 50 and 100 ppm. The resulting R2 values are 0.676, 0.651 and 0.586, respectively. For the joint predictions, in our case we still obtained better R^2 score in spite of adding noise. The R^2 was between 0.73 - 0.74 on addition of noise.

2.1.2 Curated dataset using ExoCTK Generic Grid

We have also generated data using Web scrapping method from a already published grid ExoCTK [12] which is a hot Jupiter specific grid. The data is not public we will check on its validity. It will be released after the accuracy of dataset is confirmed. We generated 112640 dataset, each containing 5000 features and 8 parameters(Temperature (T), gravity (g), radius of planet (R_p), radius of star (R_s), metallicity, C/O ratio, cloud and haze).Training and testing dataset are made by randomly dividing in the ratio of 4:1. The data values to be generated were decided by the paper [13]. The data is in the .csv format.A .csv file contains whole set of data. In the future works we plan to use the dataset to improving the model and making a more robust model. The spectra generated is shown in Figure 4.

2.2 Training

2.2.1 For HELA Dataset

We have tried various regression methods and we chose the models that gave better R^2 scores. We have performed training using support vector regressor (SVR), XGBoost regressor , Gradient Boosting, k-nearest neighbour to check the R^2 scores. Thus we chose Random Forest , KNN and Gradient Boosting as our we have used an **Ensemble Learning Approach** as our machine learning technique. We have made the Ensemble of Random Forest , Gradient Boosting and K-Nearest

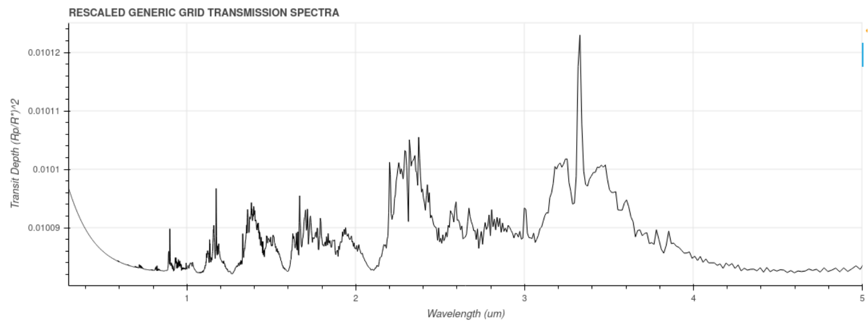

Figure 4: *Spectra as it looks like (ExoCtk Generic Grid). X axis contains 5000 values of wavelength, y axis is transit depth* [12].

Neighbours which were used as base model accompanied by a Ridge Regressor which was used as a meta model to predict the best output from the Ensemble learning.
Random Forests can capture non-linear relationships and handle high-dimensional data, kNN can identify local patterns and similar atmospheres, and Gradient Boosting can effectively model complex relationships and automatically select relevant features. The Ridge regressor is used to get the best prediction considering all the base models.The meta-model (Ridge Regression) acts as a regularizer, helping to prevent over fitting and improving the generalization performance of the ensemble model. We have used stackingregressor() in multipleoutputregressor() which is used as it is a multivariate problem. For each of the base models the hyper parameter tuning was done, for Random forest the best parameter that is n-estimator = 100, and KNN the hyper parameter tuning was done using Grid search and the value of k was found out to be 11, and metric was euclidean for the HELA dataset. The training was done without adding noise and also after adding 50 ppm ,100 ppm noise as mentioned earlier , testing has also been done by adding noise 100ppm , the noise was in the form of a Gaussian noise/ppm.

2.2.2 How the stacking works

We have used stackingregressor(), a module of sci-kit learn. The stacking regressor uses k fold cross validation method, by default its set to k= 5.The training set we have used (80,000 training points) thus is divided into k folds , k-1 batches of input and kth validation set. For each of the base model prediction is done using the validation set, thus we will have 3 different predictions from 3 different base models, these prediction matrix or feature matrix becomes the training set of our base model that is ridge regressor. Ridge regressor is trained on the prediction by base models and the corresponding true value (from the validation set). Once training is done.The model is ready for retrieval (after testing). When a observational spectrum is sent in the input then all the 3 base models give the prediction output and based on the output received from base models, ridge regressor gives a final prediction based on its training.

2.3 Testing Results

For testing accuracy R^2 score was taken as the accuracy parameter, the main reason for this was that most of the papers have used $R^2 score$ so it becomes easy for us to compare our model with other papers. The R^2 Score for HELA as mentioned in the paper was 0.676 for 50ppm noise. Surprisingly our model outperforms HELA. Our model R^2 **score was calculated as to be 0.751** for 50ppm noise and also better for other noise modulations. HELA Dataset is specific for Hot Jupiter type of exoplanet obtained from WFC3, for this data set our model gives the highest accuracy from who so ever have used HELA data set for training (Madhusudhan ,ExoCNN [14]). Rest of the paper have worked on enhancing Neural network methods with different type of dataset which are not publicly available.Our main focus was improving on random forest. In the future works we would be implementing other methods as well. To validate whether the prediction is accurate or not we wanted to regenerate the spectra with the predicted labels and compare it with the spectra from which label has been predicted, but HELA data set has only 13 values of transit depth as they must have done feature extraction thus regenerating spectra using Hela dataset is not possible that's why we generated dataset using ExoCtk

Table 1: R^2 *score of different algorithms run using Hela dataset, To see how different models perform and we chose the best models for our ensemble stacking model.*

ML Baseline algorithms	R^2 score
SVR	0.570
XGBoost	0.732
Random Forest	0.747
Gradient Boosting	0.660
kNN	0.744

Generic grid and run on the model we used, it has 5000 fetaure and 8 labels and it doesn't give a satisfying R^2 score, so we used ExoCTK dataset we generated to run different models (RF ,KNN ,CNN) without feature engineering . We see a scope of improvement of R^2 score. Figure13 shows the individual parameter R^2 score of parameters for the our Ensemble stacking model and the values are comparable to HELA as shown in Figure1.

2.3.1 Comparison between Different ML models

Besides Running the Ensemble learning model. To compare how Different Models work, we ran different algorithms including Support vector Machine (SVR) , Random Forest , XGBoost , K-nearest Neighbour(KNN) on HELA dataset.The R^2 values of the models run are shown in the Table 1. Different models were run to obtain the R^2 Score. The individual parameter prediction accuracy and corner plots generated are shown. Figure5 shows the real vs predicted plot of individual parameter when run on SVR. Figure6 shows the corner plot which shows the predicted values. Figure7 shows the real vs predicted plot of individual parameter when run on KNN, Figure8 shows the corner plot of retrieved parameters using KNN. Figure9 shows the real vs predicted plot of individual parameter when run using Random forest, Figure10 shows the corner plot of retrieved parameters using Random Forest. Figure11 shows the real vs predicted plot of individual parameter when run using XGBoost, Figure12 shows the corner plot of retrieved parameters using XGBoost.*Please find all the plots at the end of the report*.

3 Conclusion:

Atmospheric Retrieval of exoplanet is a method for characterisation of atmosphere of exoplanet which follows different chemistry and T-P profile. Machine learning can replace the traditionally used methods like statistical Bayesian inference methods (MCMC, nested sampling) which is based on finding likelihood functionality. Machine learning techniques can beat the Non ML techniques both in terms of accuracy and computational cost making atmospheric retrieval run on the fly. In our report, we have created an ensemble-learning of Random forest, Gradient boosting and kNN, which is evaluated using stacking regressor (cross validation) and the final prediction was done using ridge regressor. The model was tested on HELA Data set and the R^2 value obtained was 0.7512 which was found better than HELA and gave comparable individual R^2 score. **Not only that our model gives the highest R^2 score for that dataset till now**.
We also would be checking if the R^2 score is communicating the accuracy in the right way. We also generated around 112640 data from ExoCTK. since the model contains 5000 feature, it becomes necessary to opt for a feature engineering method, the same is under process as of now. We have also run some models like CNN, KNN, Gradient Boosting , Random Forest without hyper parameter tuning and evaluated the R^2 score, there is scope for improvement of R^2 score which we we would be doing in the future.

3.1 Limitation of Machine Learning method:

- **Generalisability** — Any changes to the underlying model,spectral range or resolution, will hinder the model's performance, and in some cases, will require a full re computation of the training data from scratch and retraining the model.

- **Lack of Bayesian framework** — Most contemporary retrievals aim to map the Bayesian posterior distribution. In contrast, most ML models applied in the field of atmospheric characterisation are formulated to perform maximum likelihood estimation. The difference between these two objectives presents an obstacle when trying to compare outputs from the two methodologies.

4 Future plans :

We will now be working with the ExoCtk data we generated. Data so generated would require the model which would cater the complexity of the data, once the R^2 is satisfactory we can regenerate spectra with the labels predicted and see whether we generate the same spectra which was used for prediction this will validate our accuracy and prediction.
We are working for integrating ML- Based Retrieval with the in house Retrieval Algorithm under the exoplanet and planetary formation group (PI: Dr Liton Majumdar) which contains a Spectra generator, opacity calculator and a Retrieval module.The Spectra generator would have information about different type of exoplanet and the data we would generate would be a wide range dataset, which is not present in the community. Our machine learning algorithm would take into account for both accuracy and computational time. A comparative analysis would also be given between different ML models.

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] https://www.space.com/17738-exoplanets.html

[6] Madhusudan,2018;https://arxiv.org/pdf/1808.04824.pdf

[7] POSEIDON; https://github.com/MartianColonist/POSEIDON

[8] ExoGAN :https://github.com/ucl-exoplanets/ExoGAN$_public$

[9] Plan net ; https://github.com/exoml/plan-net

[10] exoCNN; https://gitlab.astro.rug.nl/ardevol/exocnn

[11] HELA github repository; https://github.com/exoclime/HELA

[12] ExoCTK data scraping; https://exoctk.stsci.edu/

[13] Goyal, J. M., Wakeford, H. R., Mayne, N. J., Lewis, N. K., Drummond, B., & Sing, D. K. (2018, November 3). Fully scalable forward model grid of exoplanet transmission spectra. Monthly Notices of the Royal Astronomical Society, 482(4), 4503–4513.

[14] Martínez, F. A., Min, M., Kamp, I., & Palmer, P. I. (2022). Convolutional neural networks as an alternative to Bayesian retrievals for interpreting exoplanet transmission spectra. Astronomy & Astrophysics, 662, A108. https://doi.org/10.1051/0004-6361/202142976

[15] **All the codes used for our project can be found here**.

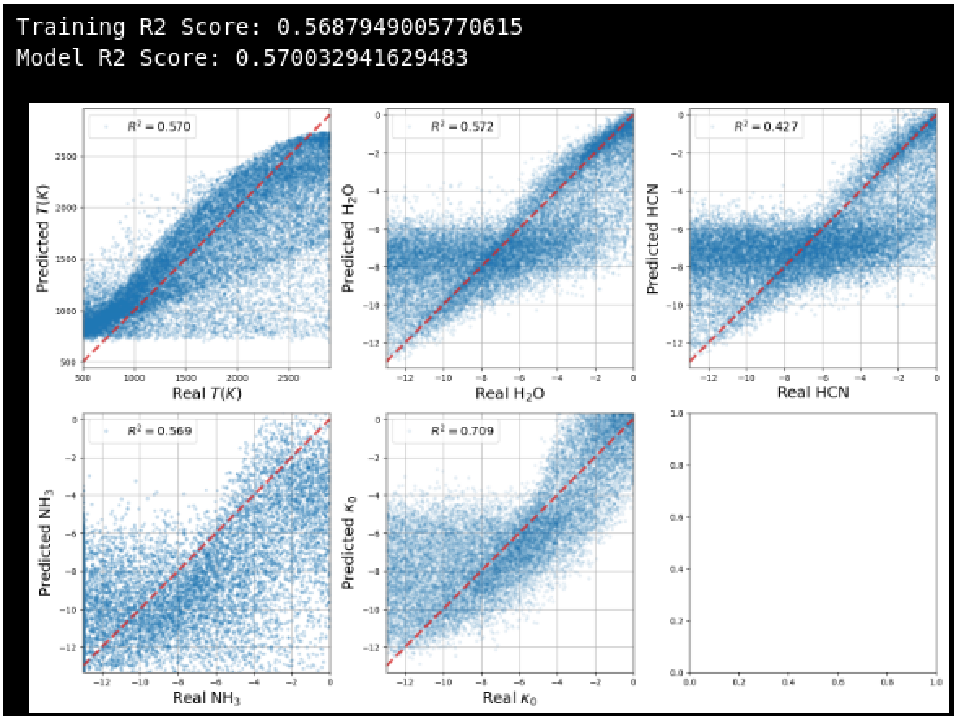

Figure 5: Individual parameter RvP plot for SVR using HELA dataset, R^2 for T(K) = 0.570 , H_2o = 0.572 , HCN = 0.427 , NH_3 = 0.569 , k_0 = 0.709, on Top : Model R^2 = 0.570

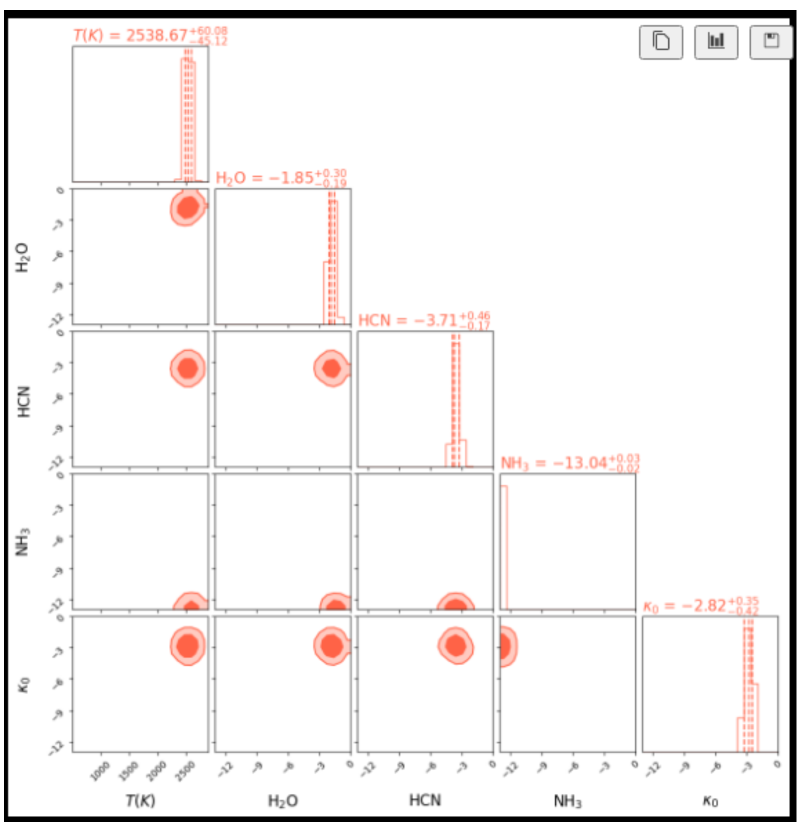

Figure 6: Corner plot for SVR

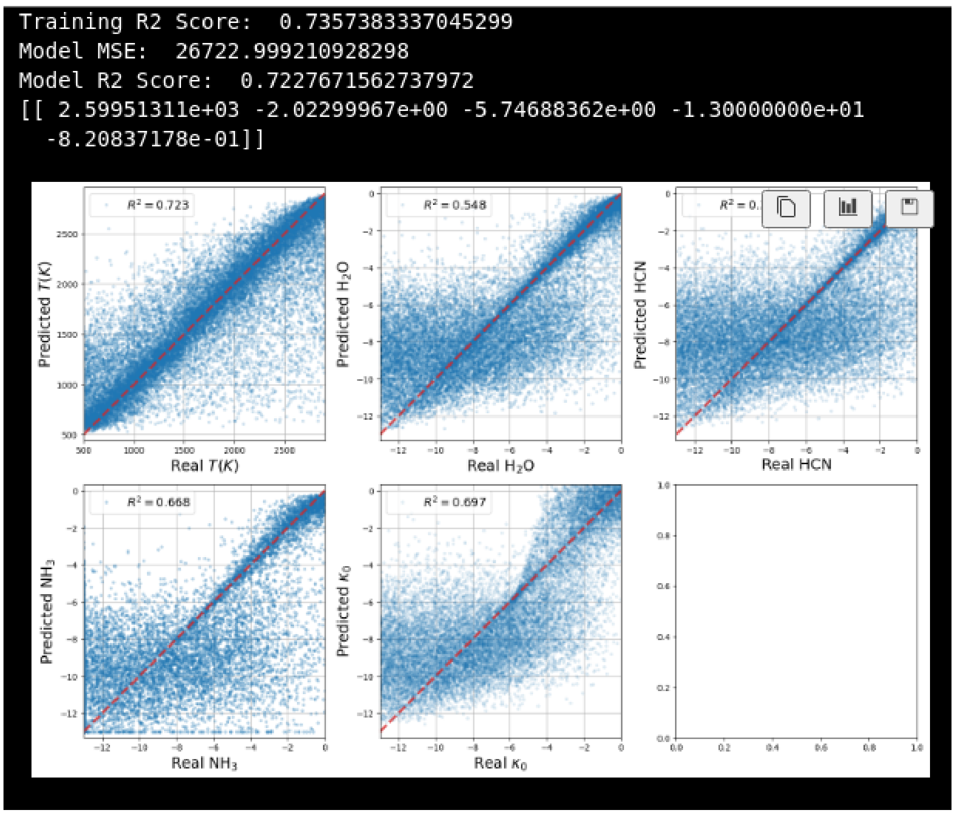

Figure 7: Individual parameter RvP plot for KNN using HELA dataset,R^2 for T(K) = 0.723 , H_2o = 0.548 , HCN = 0.502 , NH_3 = 0.608 , k_0 = 0.697, on Top : Model R^2 = 0.722

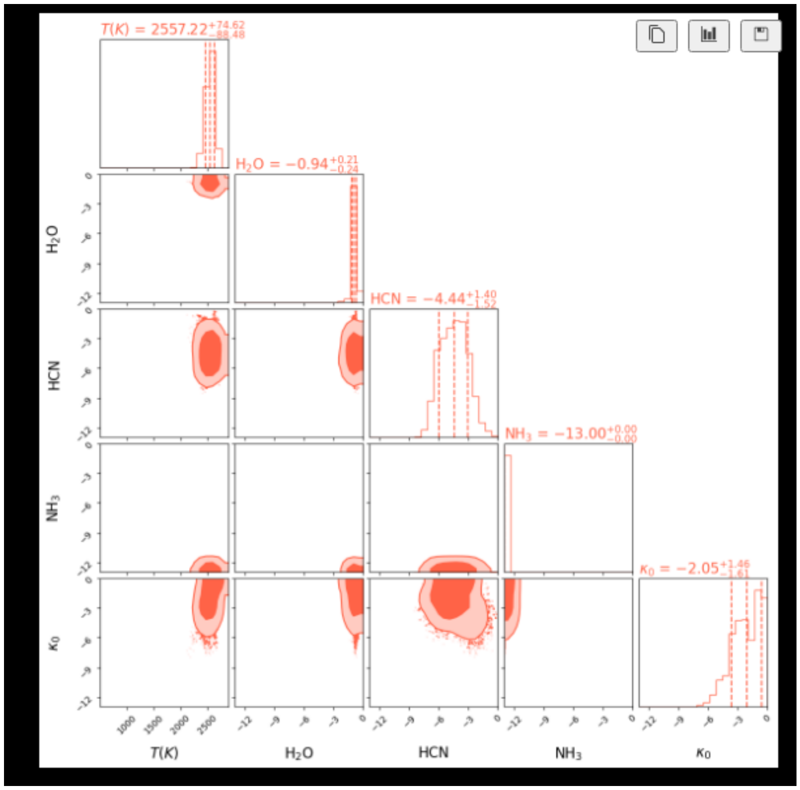

Figure 8: Corner plot for KNN

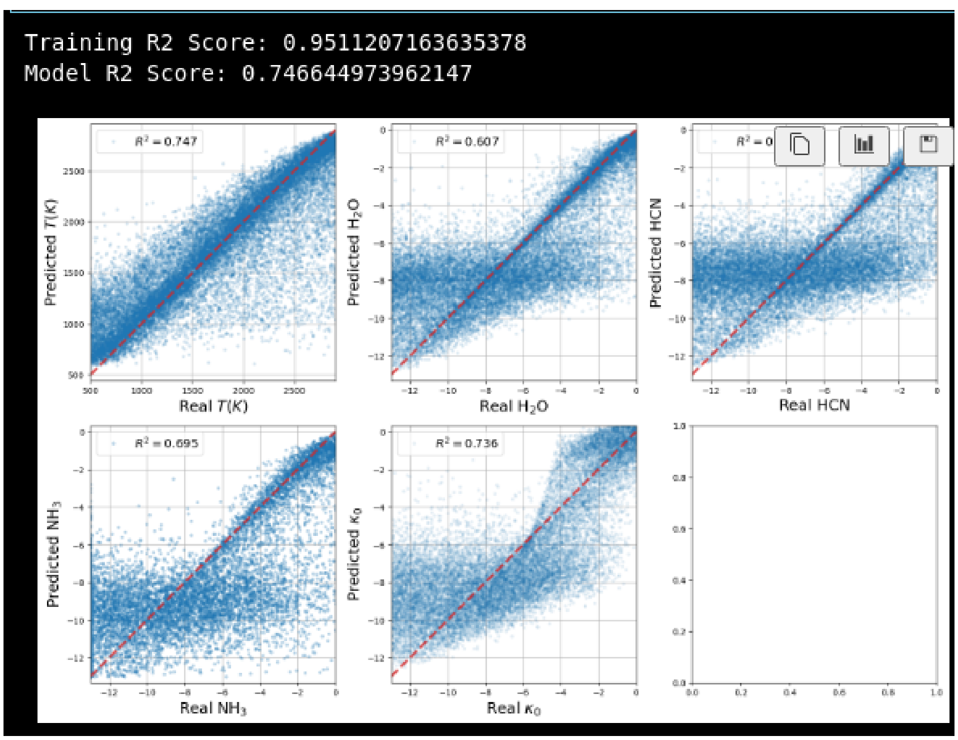

Figure 9: Individual parameter RvP plot for Random Forest using HELA dataset,R^2 for T(K) = 0.747 , H_2o = 0.607 , HCN = 0.497 , NH_3 = 0.695 , k_0 = 0.736, on Top : Model R^2 = 0.746

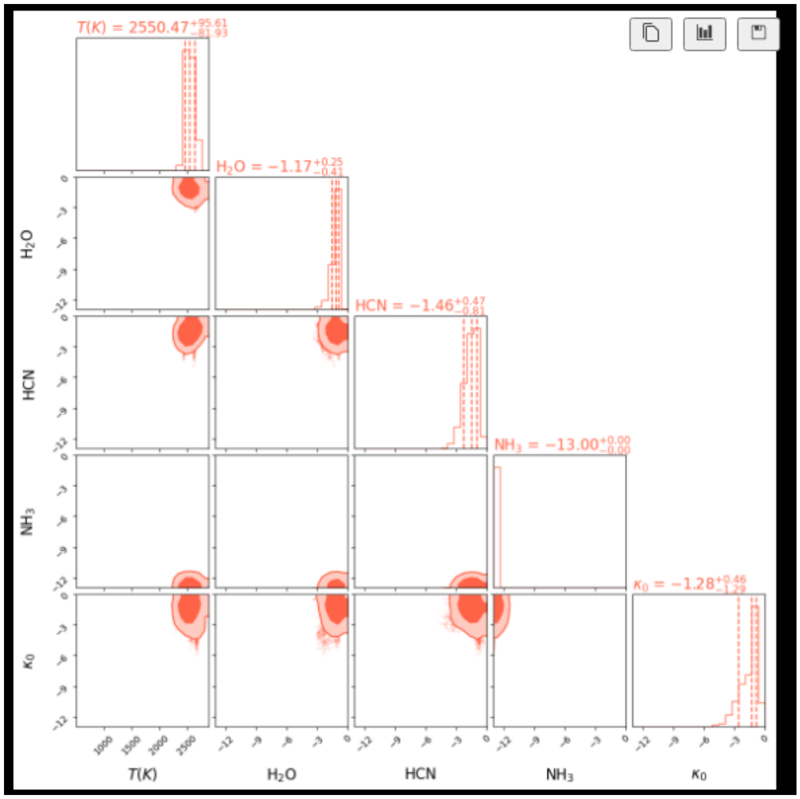

Figure 10: Corner plot for Random Forest

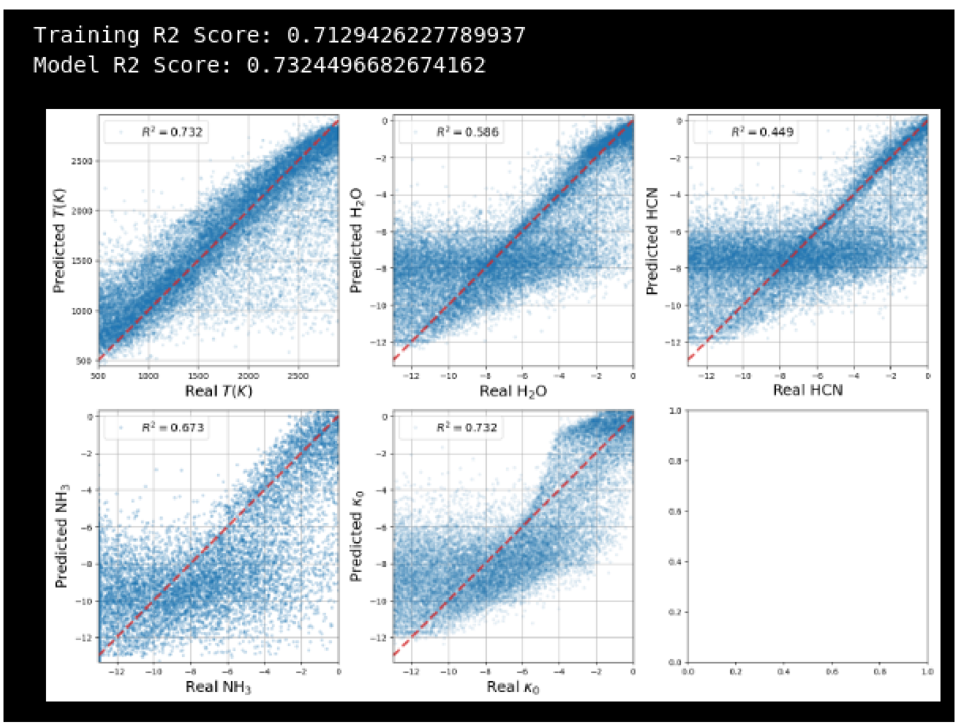

Figure 11: Individual parameter RvP plot for XGBoost using HELA dataset,R^2 for T(K) = 0.732 , H_2o = 0.586 , HCN = 0.449 , NH_3 = 0.673 , k_0 = 0.732, on Top : Model R^2 = 0.732

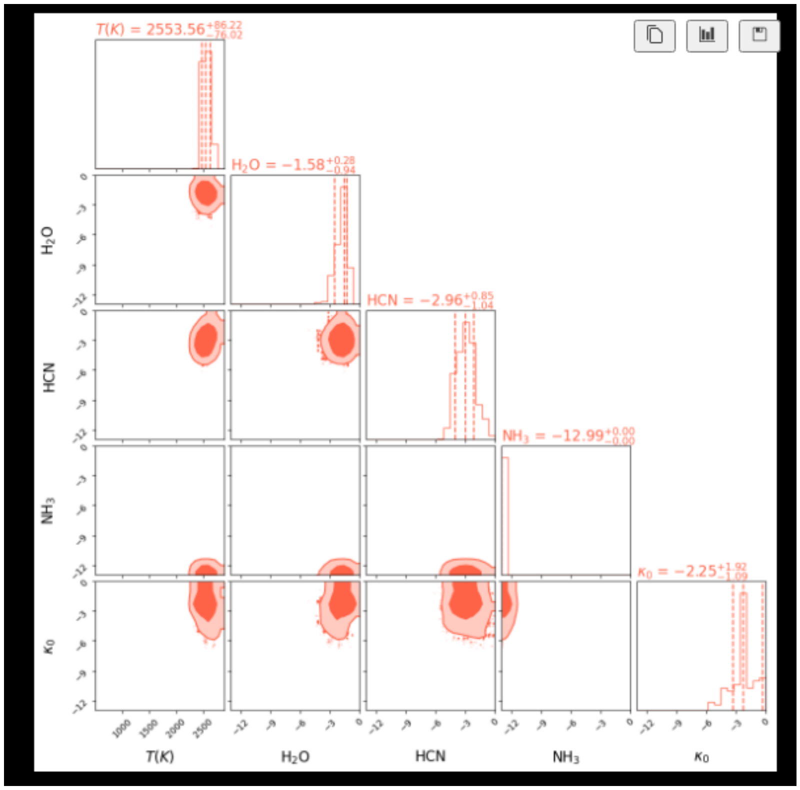

Figure 12: Corner Plot for XGBoost

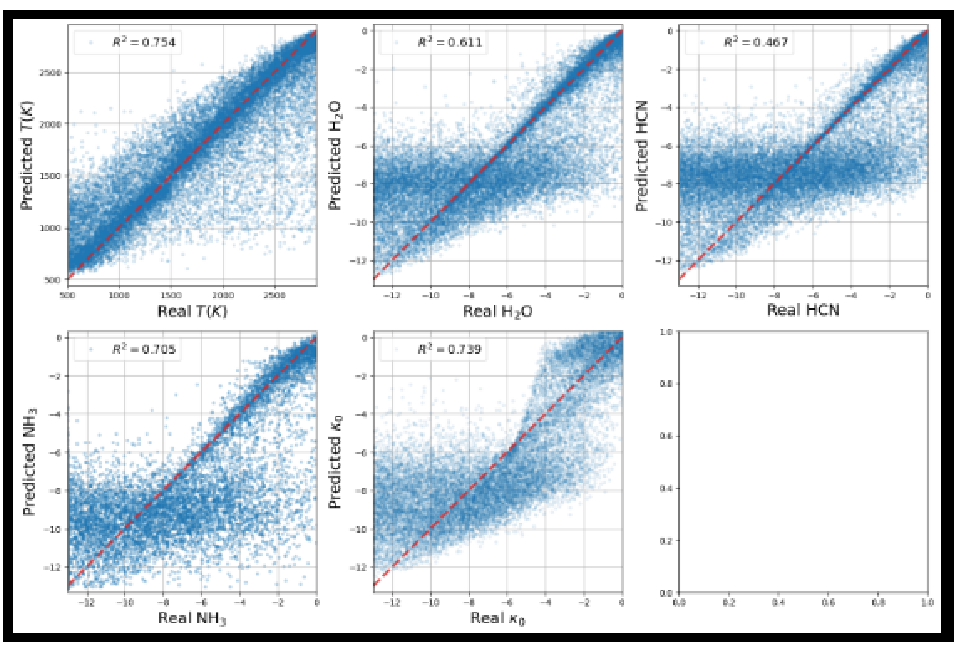

Figure 13: Individual parameter RvP plot for Ensemble Stacking (MLARE) using HELA dataset,R^2 for T(K) = 0.754 , H_2o = 0.611 , HCN = 0.467 , NH_3 = 0.705 , k_0 = 0.739, on Top : Model R^2 = 0.752

Tasneem Basra Khan Paper Check

By Tasneem Basra Khan

WORD COUNT	2691	TIME SUBMITTED	19-APR-2024 08:21PM
		PAPER ID	108383945

MLARE : ML Driven Atmospheric Retrieval of Exoplanet

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

Abstract

The paper illustrates the use of Machine Learning in Atmospheric Retrieval of Exoplanets focusing mainly on Hot Jupiter transmission spectra. We introduce here the Ensemble learning of Random forest , Gradient Boosting and K-Nearest Neighbours and bench marked with other Pre-existing machine learning especially Random Forest induced Atmospheric retrieval. To check the prediction. R^2 score was found out to be 0.752 with added 50ppm and 100ppm noise and bench marked with HELA in which the R^2 was 0.676, 0.651 and 0.586 for noise floors of 10, 50 and 100 ppm. Besides that the Data curation was also done using ExoCTK Generic Grid (generation of 112640 data , 5000 transit depth values with 8 lables). In the Future works we will be working with the synthetic data generated using ExoCTK to improve on the Retrieval model in terms of accuracy and computation.

1 Introduction

1.0.1 What are Exoplanets

Exoplanets are extra solar planets that orbit stars other than the Sun outside the solar system.(1). We have more than 5000 exoplanets discovered so far. 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 physiochemical 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. Based on the composition(Mixing Ratios, equilibrium or non equilibrium chemistry). The Exoplanets can be classified as Hot Jupiter , Rocky/terrestrial , Sub-neptune etc. In this paper we mainly consider the data of hot Jupiters (like WASP39b), these are closer to the star and have masses similar or greater to the Jupiter, we also have rocky exoplanet. The studies on these exoplanet is also based on which kind of resolution telescope we are using. A higher resolution telescope gives a better wavelength scale and we can get a wider information about the atmosphere. We will mainly consider JWST data and WFC3 (Wide field Camera).

1.0.2 Atmospheric Retrieval a review

Atmospheric Retrieval is the technique which helps us to know about the constituents of chemicals or other properties hidden in the intricacies of exoplanetary atmospheres. The atmosphere may contain molecule and the absorption of light due to that molecule defines the opacity and reflection of light due to molecule is called the albedo moreover its the transmission, absorption or refelction of light from the exoplanet that reaches the telescope is what used to predict the molecules and properties. The relationship between different properties is given by the Radiative transfer equation. The atmosphere can be divided into different layers to get the Temperature-Pressure profile across different layer and different phenomenon like Collision induced absorption, pressure broadening come into play, Based on the dynamics of atmosphere we can have equilibrium chemistry or have non equilibrium chemistry which is mainly due to convection or wind in the atmosphere and thats why we will be characterising exoplanets(The properties like metallicity, C/O ratio, cloud, haze and chemical abundances like Water, Ammonnia etc.) to study the atmosphere of the exoplanet to know about its component and the phenomenon that drives their abundances. Our main focus would be finding the chemical abundance, Effective temperature that is the average temperature of the exoplanet, Metallicity that how much metal like elements are there, C/O ratio that is how much of carbon oxygen is present, constant cloud opacity which will tell us more about the opacity of transmission.

1.0.3 Some Previous works

Various traditional methods have been used so far including Bayesian inference methods such as MCMC, nested sampling(Pymultinest , Ultranest etc). Pymultinest gives a bencmark accurate prediction and is used in various Atmospheric Retreival models like Poseidon., Picaso etc which do not use machine learning. Here we would like to introduce Machine learning for increasing the accuracy and bringing up efficient computational power. Many Atmospheric retrieval methods using Machine learning do exist in the community,but are

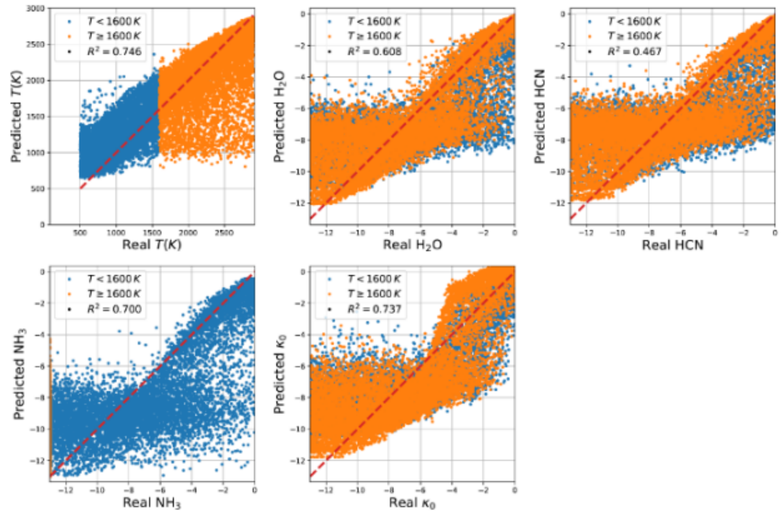

Figure 1: R^2 prediction of Individual parameter (HELA)

not widely used. The first machine learning technique which came used Random forest method and the Model was called HELA. Márquez-Neila, P., Fisher, C., Sznitman, R., Heng, K. (2018, June 25). Supervised machine learning for analysing spectra of exoplanetary atmospheres. Nature Astronomy, 2(9), 719–724. The Prediction is specific to the type of exoplanet, it can be a Hot Jupiter or terrestial exoplanet and it can be either transmission spectra or emmision or refelction spectra. Our main focus would be Prediction of hot jupiters on the transmission spectra data. There are few Machine Learning models using that category namely HELA ,ExoGAN , Plan net , Fisher , Madhusudhan , ExoCNN , VI retrieval. HELA as said earlier uses Random Forest, ExoGAN uses Generative Adversial Network , Plan net uses Bayesian Neural network, Fisher madhusudan have worked on Random forest , ExoCNN is a Convulational neural network technique and we also variational inference as well. We will working on JWST/WFC3 Transmission spectra for hot jupiters and our main focus would be benchmarking the previous technique especially the technique which uses random forest like HELA and madhusudhan. We have generated the corner plots after running the already existing ones namely of POSEIDON and HELA as it is shown in the figure, and we expect the comparable R^2 score.

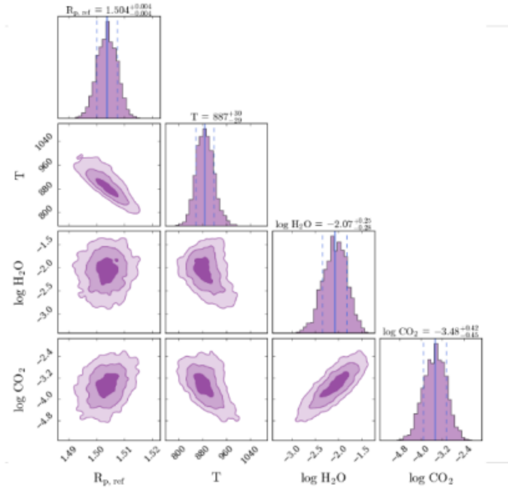

Figure 2: Corner plot for POSEIDON Retrieval

2 Methodology

2.1 Dataset Used for Training

2.1.1 HELA dataset: WFC3 Transmission Spectra

In this paper we are first using data set already available in the community and many other papers have also used the same data set(HELA data set). The training set consists of 80,000 synthetic WFC3 transmission spectra, each described by 5 parameters: the temperature (T), volume mixing ratios (relative abundances by number) of water (X_{H_2O}), ammonia (X_{NH_3}) and hydrogen cyanide (X_{HCN}), and a constant cloud opacity (κ_0). The training set resides in a 13-dimensional space, where each dimension corresponds to a wavelength bin. Along the axis of each dimension is a continuous range of values of the transit radii. The testing dataset has 20,000 synthetic spectra, arranged in similar fashion. HELA dataset is a synthetically generated WFC3 dataset. The data set can be found in the HELA github repositary, before training we have used standardscaler() besides that robustscaler() and MinMaxscaler() was also used to scale the data but gave same results and also added noise as 50ppm training and 100ppm testing to avoid overfitting. In the HELA paper noise was assumed to be a Gaussian uncertainty on the transit depths with full widths at half-maximum of 10, 50 and 100 ppm which represent ideal, typical and easily attainable conditions. As a further check they assumed noise floors of 10, 50 and 100 ppm. The resulting R2 values are 0.676, 0.651 and 0.586, respectively, for the joint predictions, in our case we still obtained the better R^2 score inspite of adding noise. The R^2 was between 0.73 - 0.74 on addition of noise.

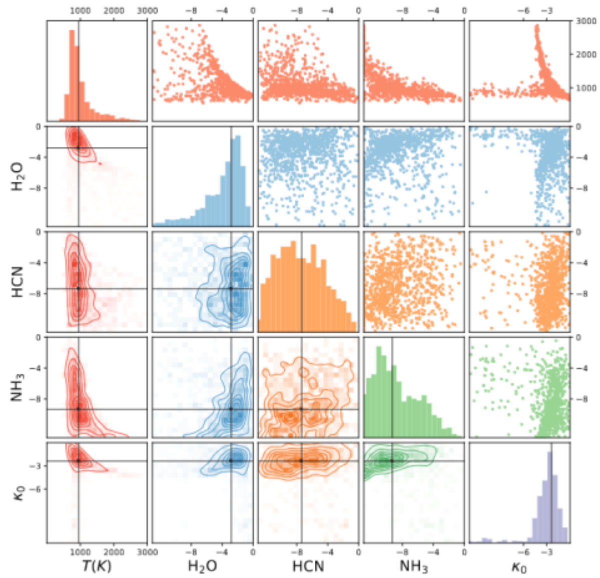

Figure 3: Corner plot for HELA Retrieval

2.1.2 Curated dataset using ExoCTK Generic Grid

We have also generated data using Webscrapping method from a already published grid ExoCTK which is a hot Jupiter specific grid. The data is not public we will check on its it will be released after the accuracy of dataset is confirmed. We generated 112640 dataset, each containing 5000 features and 8 parameters(Temperature (T), gravity (g), radius of planet (R_p), radius of star (R_s), metallicity, C/O ratio, cloud and haze).Training and testing dataset are made by randomly dividing in the ratio of 4:1. The data values to be generated were decided by the paper Goyal, J. M., Wakeford, H. R., Mayne, N. J., Lewis, N. K., Drummond, B., Sing, D. K. (2018, November 3). Fully scalable forward model grid of exoplanet transmission spectra. Monthly Notices of the Royal Astronomical Society, 482(4), 4503–4513.. The data is in the .csv format.A .Csv file contains whole set of data.This dataset generated is also scaled and feature engineering has been done on the data to get the optimum results. In the future works we plan to use the dataset to improving the and making a more robust model.

2.2 Training

2.2.1 For HELA Dataset

Our main focus has been working on Random Forest. We have tried various regression method and we chose the models that already give a better R^2 score. We have performed training using support vector regressor (SVR), XGBoost

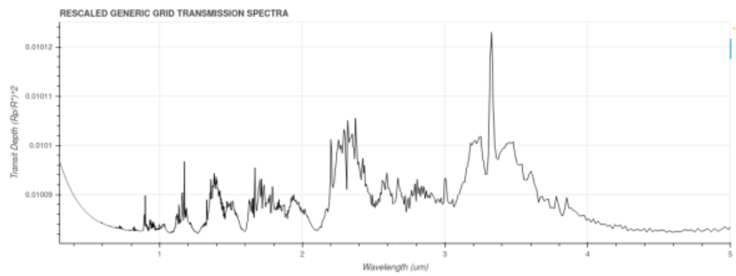

Figure 4: Spectra as it looks like (ExoCtk Generic Grid)

regressor , Gradient Boosting, k-nearest neighbour to check the R^2 score. So we have used an **Ensemble Learning Approach** as our machine learning technique. We have made the Ensemble of Random Forest , Gradient Boosting and K-Nearest Neighbours which were used as base model accompanied by a Ridge Regressor which was used as a meta model to predict the best output from the Ensemble learning.
Random Forests can capture non-linear relationships and handle high-dimensional data, kNN can identify local patterns and similar atmospheres, and Gradient Boosting can effectively model complex relationships and automatically select relevant features. stackingregressor() from sklearn has been used for the implementation of ensemble learning, it uses cross validation. Cross-validation is done by splitting the training data into multiple folds, training the base models on some folds, and using them to make predictions on the remaining folds improving on each batch. Then a Ridge regressor is used to get the best prediction considering all the base models.The meta-model (Ridge Regression) acts as a regularizer, helping to prevent overfitting and improving the generalization performance of the ensemble model. multipleoutputregressor() is used as it is a multivariate problem. For each of the base models the hyperparameter tuning was done, for Random forest the best parameter that is n-estimator = 100, and KNN the hyperpaarmeter tuning was done using Grid search and the value of k was found out to be 11, and metric was euclidean for the HELA dataset.
The training was done without adding noise and also after adding 50 ppm ,100 ppm noise as mentioned earlier , testing has also been done by adding noise 100ppm , the noise was in the form of a Gaussian noise/ppm.

2.3 Testing Results

The R^2 Score for HELA as mentioned in the paper was 0.676 for 50ppm noise. Surprisingly our model outperforms HELA. Our model R^2 **score was calculated as to be 0.751** for 50ppm noise. The R^2 values of the models run are shown in the table below:

2.3.1 Comparison between Different ML models

Besides Running the Ensemble learning model. To compare how Different Models run on HELA dataset, different models were run to obtain the R^2 Score and

ML Baseline algorithms	R^2 score
SVR	0.570
XGBoost	0.732
Random Forest	0.747
Gradient Boosting	0.660
kNN	0.744

corner plots and the results are shown below.

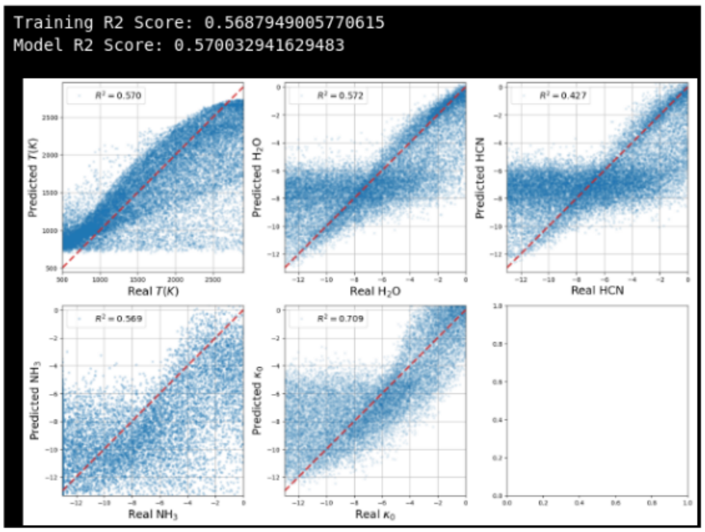

Figure 5: RvP plot for SVR machine learning

3 Conlusion:

Atmospheric Retrieval of exoplanet is a method for characterisation of atmosphere of exoplanet which follows different chemistry and TP profile. Machine learning can replace the traditional used method like statistical bayesian inference method which is based on finding liklihood functionality. Machine learning techniques can beat the Non ML techniques both in terms of accuracy and computational cost making atmospheric retrieval run on the fly. In our Paper we have created an ensemble learning of Random forest, Gradient boosting and kNN, which is evaluated using stacking regressor (cross validation) and the final prediction was done using ridge regressor. The model was tested on Hela Data set and the R^2 value obtained was 7.512 which was found better than HELA and gave comparable individual R^2 score. Not only that our model gives the highest R^2 score on the dataset till now.
We also would be checking if the R^2 score is communicating the accuracy in

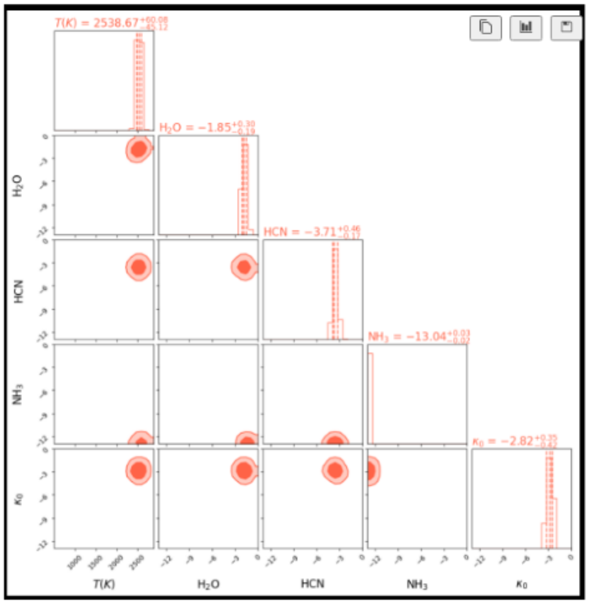

Figure 6: Corner plot for SVR

the right way. We also generated around 112640 data from ExoCTK. since the model contains 5000 fetaure, it becomes necessary to opt for a feature engineering method, the same is under process as of now. We have also run some models like CNN, KNN, Gradient Boosting , Random Forest without hyperparameter tuning and evaluated the R^2 score, there is scope for improvement of R^2 score which we we would be doing in the future.

3.1 Limitation of Machine Learning method:

- **Generalisability** —Any changes to the underlying model,spectral range or resolution, will hinder the model's performance, and in some cases, will require a full recomputation of the training data from scratch and retraining the model.
- Lack of Bayesian framework Most contemporary retrievals aim to map the Bayesian posterior distribution. In contrast, most ML models applied in the field of atmospheric characterisation are formulated to perform maximum likelihood estimation. The difference between these two objectives presents an obstacle when trying to compare outputs from the two methodologies.

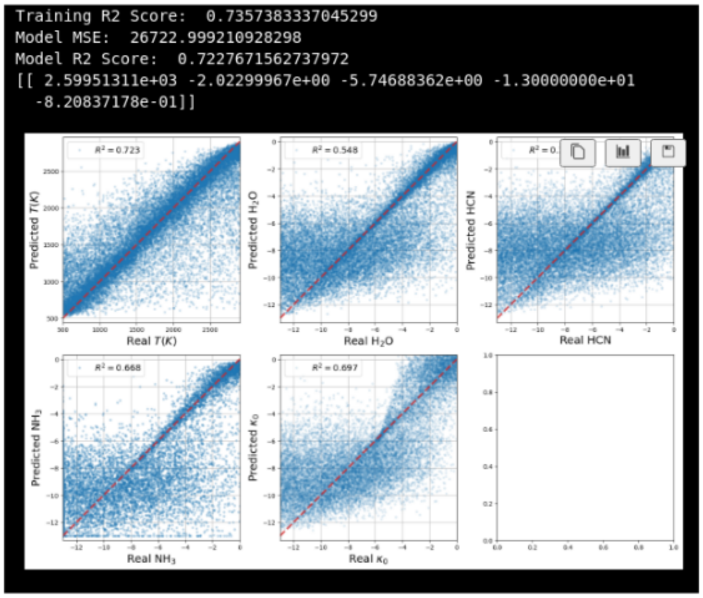

Figure 7: RvP for KNN

4 Future plans :

We will now be working with the ExoCtk data we generated, Data so generated would require the model which would cater the complexity of the data.
We are working for integrating ML- Based Retrieval with the in house Retrieval Algorithm under the exoplanet and planetary formation group (PI: Dr Liton Majumdar) which contains a Spectra generator, opacity calculator and a Retrieval module.The Data which would be a wide range would be taken from the Spectra generator.Our machine learning algorithm would take into account for both accuracy and computational time. A comparative analysis would also be given between different ML models.

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.

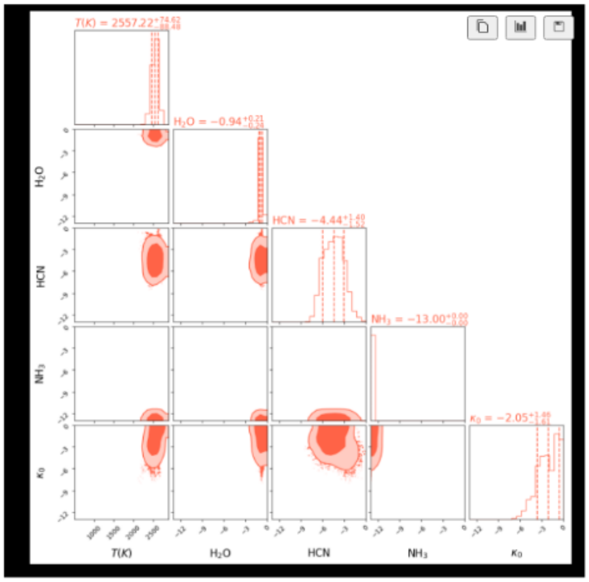

Figure 8: Corner plot for KNN

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. https://www.space.com/17738-exoplanets.html

6. Madhusudan,2018;https://arxiv.org/pdf/1808.04824.pdf

7. POSEIDON; https://github.com/MartianColonist/POSEIDON

8. ExoGAN :https://github.com/ucl-exoplanets/ExoGAN$_{p}ublic$

9. Plan net ; https://github.com/exoml/plan-net

10. exoCNN; https://gitlab.astro.rug.nl/ardevol/exocnn

11. HELA github repository; https://github.com/exoclime/HELA

12. ExoCTK data scraping; https://exoctk.stsci.edu/

13. Goyal, J. M., Wakeford, H. R., Mayne, N. J., Lewis, N. K., Drummond, B., Sing, D. K. (2018, November 3). Fully scalable forward model grid of exoplanet transmission spectra. Monthly Notices of the Royal Astronomical Society, 482(4), 4503–4513.

14. All the codes used for training and testing (in HELA) can be found here.

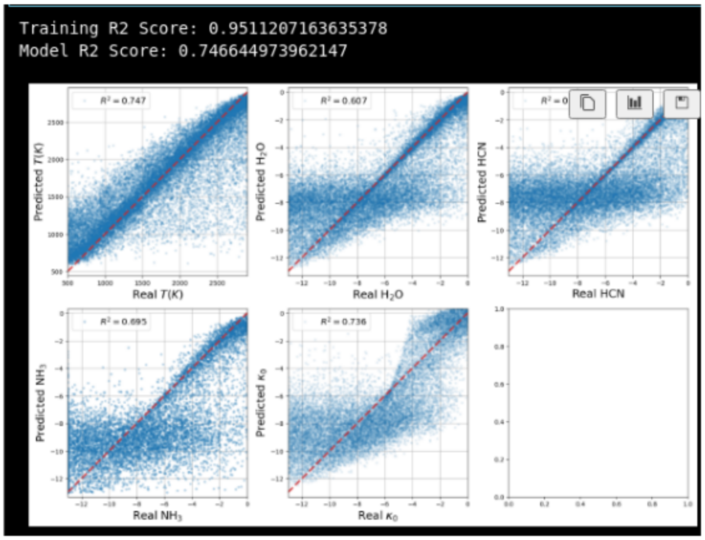

Figure 9: RvP for Random Forest

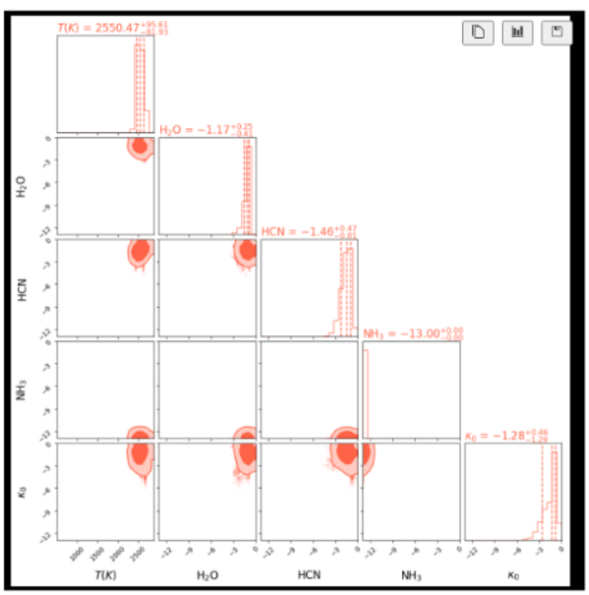

Figure 10: Corner plot for Random Forest

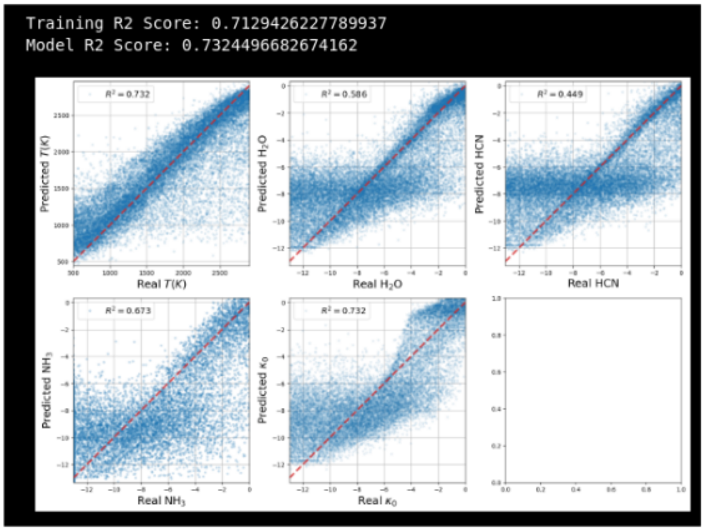

Figure 11: RvP plot for XGBoost

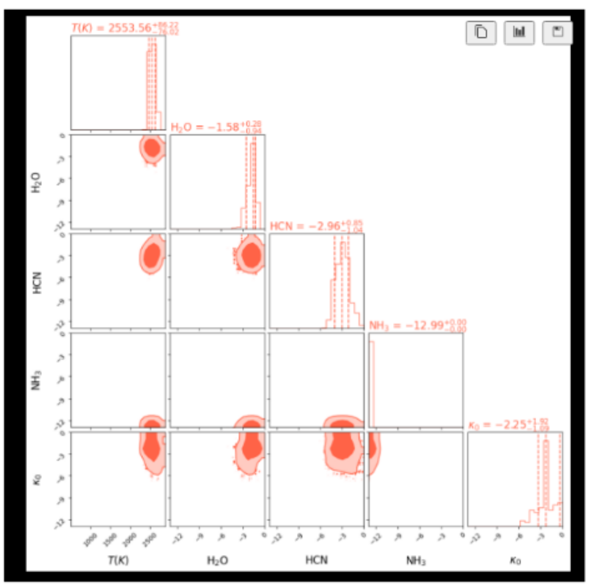

Figure 12: Corner Plot for XGBoost

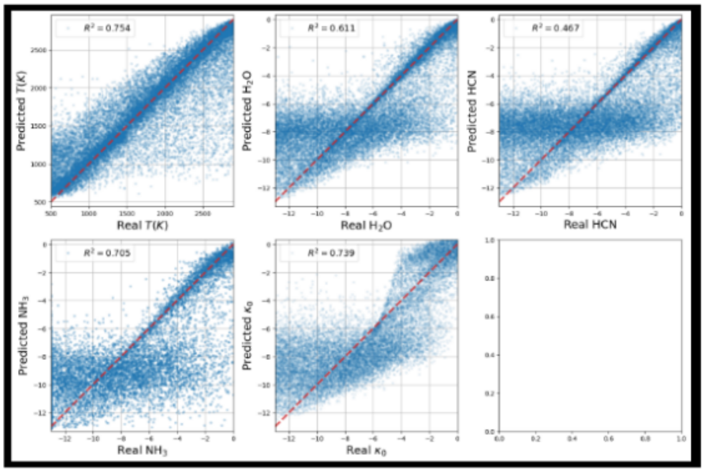

Figure 13: RvP plot for Stacking

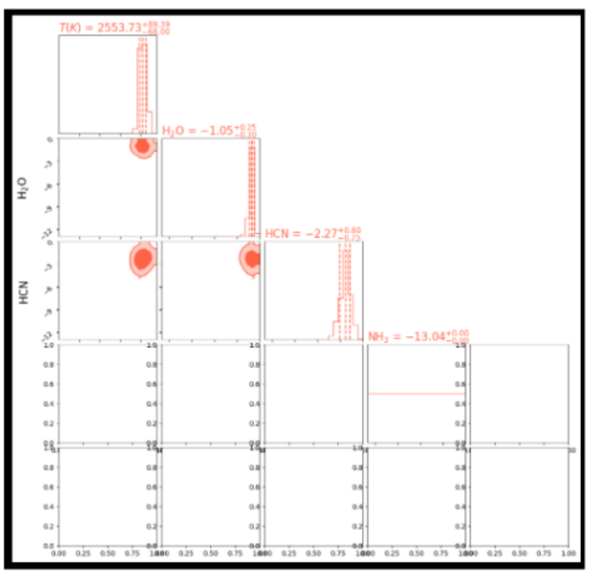

Figure 14: Corner Plot for Stacking

Tasneem Basra Khan Paper Check

ORIGINALITY REPORT

13%
SIMILARITY INDEX

PRIMARY SOURCES

1	www.arxiv-vanity.com Internet	105 words — 4%
2	iopscience.iop.org Internet	83 words — 3%
3	edoc.ub.uni-muenchen.de Internet	36 words — 1%
4	boristheses.unibe.ch Internet	31 words — 1%
5	arxiv.org Internet	25 words — 1%
6	Adam D. Cobb, Michael D. Himes, Frank Soboczenski, Simone Zorzan et al. "An Ensemble of Bayesian Neural Networks for Exoplanetary Atmospheric Retrieval", The Astronomical Journal, 2019 Crossref	8 words — < 1%
7	R. Udendhran, G. Yamini, N. Badrinath, J. Jegathesh Amalraj, A. Suresh. "Enhancing representational learning for cloud robotic vision through explainable fuzzy convolutional autoencoder framework", Soft Computing, 2023 Crossref	8 words — < 1%
8	oro.open.ac.uk Internet	

6 words — < 1%

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