Original Research Article

Proof-of-Concept Machine Learning Classifier for Identifying Clouds and Haze in Exoplanet Transmission Spectra

Alexander Liu

University High School, 2611 E. Matoian Way, MS/UH134, Fresno, CA 93740, United States

ABSTRACT

Traditional methods for determining whether an exoplanet atmosphere is clear, cloudy, or hazy can require substantial manual interpretation of transmission spectra. Here, a proof-of-concept machine learning (ML) classifier is developed to assess whether synthetic training data can support automated classification of atmospheric conditions in observed spectra. A dataset of 10800 synthetic spectra was generated using petitRADTRANS, spanning three classes (clear, cloudy, and hazy). The dataset was split into training, validation, and testing sets (70/15/15 percent), and a random forest classifier was trained and evaluated. The model achieved testing accuracy of approximately 99–100 percent, with cross-validated F_1 scores above 0.98 across all classes. The trained model was then applied to five observed exoplanet spectra and produced cloudy classifications with probabilities between 65 and 89 percent. Although the small sample size and synthetic training data limit generalizability, this study demonstrates the potential for ML to accelerate atmospheric characterization workflows. Future work with larger and more diverse datasets will be required to validate the method for broader scientific applications.

Keywords: exoplanet; machine learning; transit spectroscopy; transmission spectrum; random forest

INTRODUCTION

Over the past three decades, the field of exoplanet science has advanced dramatically, with nearly 6000 confirmed exoplanets that have been discovered (1-2). With the advancement of exoplanet science, exoplanet atmospheres have been of interest to researchers, specifically finding the atmospheric composition and structures of exoplanets. This is done through multiple ways, the most popular and successful being transit spectroscopy. Transit spectroscopy is a technique where scientists measure the absorption of starlight as a planet passes in front of its host star. During a transit, a small number of starlight filters through the exoplanet's atmosphere, with different gases absorbing specific wavelengths of light, creating a transmission spectrum which researchers can analyze to determine atmospheric composition (1).

Clouds and haze play distinct physical and observational roles in exoplanet atmospheres. Clouds are typically composed of larger condensate particles forming at specific pressure–temperature levels, while haze consists of small photochemically produced particles suspended at higher altitudes. Observationally, clouds generally mute or flatten absorption features across most wavelengths, whereas haze frequently introduces a Rayleigh-scattering-like slope toward shorter wavelengths. These distinctions strongly influence how atmospheric retrieval models interpret molecular abundances and scattering properties (3).

Because muted or featureless spectra obscure key

Corresponding author: Alexander Liu, E-mail: liualexander200@gmail.com.

Accepted February 27, 2026
https://doi.org/10.70251/HYJR2348.421418

diagnostic information, determining whether clouds or haze are present is a necessary step before conducting full atmospheric retrievals. However, such classification often requires manual inspection or computationally intensive retrieval modeling (1).

Machine learning (ML) offers a scalable alternative. With the increasing number of spectra expected from JWST and future missions, automated classification tools could allow rapid initial assessments of atmospheric conditions before retrieval. The aim of this study is to test whether a random forest classifier trained on idealized synthetic spectra can generalize real observational data and serve as a preliminary decision-support tool for atmospheric analysis.

METHODS AND MATERIALS

In order to create an ML model that could be used on real data, a dataset of synthetic spectra was used to train it on. The number of exoplanets with high quality known spectra, around a few dozen (2), was not enough to be used to train a ML model. This influenced the motivation to train the ML model on synthetic spectra instead, as thousands of spectra could be used to train the ML model on instead of just the few dozen real observed spectra.

PetitRADTRANS was used in order to create a dataset of 10800 synthetic spectra under three distinct classes: "clear", "cloudy" and "haze" with 3600 spectra for each class (4). For the "clear" atmosphere, first a flat spectrum was generated, then H_2O bands were added at 1.4, 1.9, 2.7, and 6.3 µm. The same was done for CH_4 bands at 1.65, 2.3, and 3.3 µm. The feature strength was randomized for each sample. For "cloudy" atmospheric conditions, it was mostly the same process. However, at the end, features were flattened by making the absorption bands less intense, simulating real atmospheric data, as clouds block light from reaching the deepest parts of the atmosphere. The process of creating "hazy" atmospheric conditions was also similar to the previous processes. However, this time, the baseline was sloped to imitate the Raleigh scattering effect of the particles in the atmosphere, with higher absorption at shorter wavelengths, tapering off as wavelength increases. This was done by randomly selecting a power-law exponent ($\approx\lambda^{-3}$ to λ^{-5}) to mimic variability in haze particle properties, producing stronger scattering at shorter wavelengths and a characteristic blueward slope. The scattering profile is normalized to isolate its shape from its amplitude, and a randomly chosen haze strength controls how strongly this continuum modifies the spectrum. The result is a realistic haze signature that can mute molecular absorption features and provides a learnable cue for classifying hazy atmospheres. Next, Gaussian noise was added to all the spectra in order to mimic JWST signal-to-noise ratios. Finally, the data was rescaled, making the average 1, so that all the spectra are on the same scale.

The next step was to train a ML classification model on synthetic data. In order to streamline this process Scikit-learn (5), a Python library with built-in metrics was used. Out of three different algorithms tested (random forest, gradient boosting, and neural network), the random forest algorithm was ultimately chosen for several factors: reduced risk of overfitting, ability to complete classification tasks, and its ability to determine which variables were most influential to determine an outcome.

Data Splitting and Validation

The 10800-spectrum dataset was randomly divided into training (70 percent), validation (15 percent), and testing (15 percent) subsets. Five-fold cross-validation was used during training to evaluate model stability and reduce overfitting.

Model Features

Each spectrum consisted of 300 wavelength points between 1 and 7 microns, giving the classifier 300 input features per sample. The model used the raw flux values directly, and no engineered features were added.

Hyperparameter Tuning

Random Forest hyperparameters were optimized using randomized search across 50 configurations, varying number of trees (200–800), maximum depth (10–50), minimum samples per split, and feature-subsampling strategy. The final chosen parameters were: 500 trees, maximum depth of 30, minimum samples per split of 2, and square-root feature sampling.

Performance Metrics

Accuracy, precision, recall, and F_1 scores were computed for the held-out test set. A confusion matrix (Figure 1) was used to examine misclassifications. The random forest classifier achieved a near-perfect classification rate on the synthetic test set. One clear sample was misclassified as cloudy. Real observed spectra were obtained from Exo.MAST (5) for five exoplanets: K2-18b, HD 209458 b, WASP-17b, WASP-12b, and HAT-P-26b.

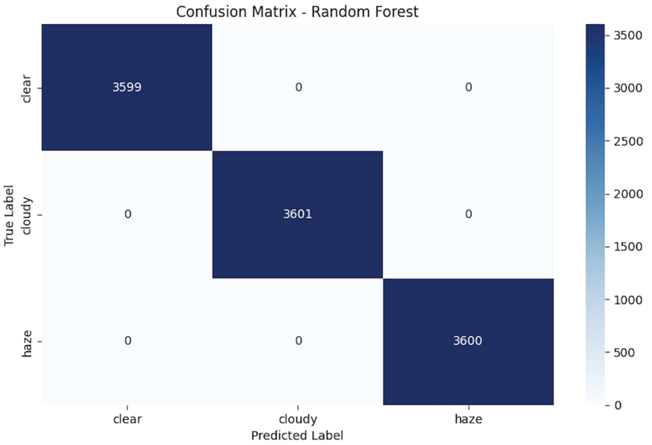

Figure 1. *Confusion matrix for the random forest model evaluated on the 15% held-out portion of the synthetic dataset. True labels are shown on the vertical axis, and predicted labels on the horizontal axis. The model exhibits high accuracy across all classes, with the only misclassification occurring between clear and cloudy spectra—reflecting the reduced feature contrast in noisy or moderately flattened spectral cases.*

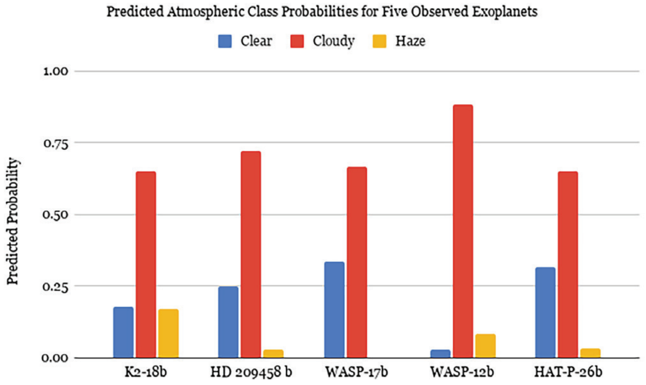

Figure 2. *Predicted class probabilities for clear, cloudy, and hazy atmospheres generated by the random forest classifier when applied to transmission spectra of five well-studied exoplanets (K2-18b, HD 209458 b, WASP-17b, WASP-12b, and HAT-P-26b). Each triplet of bars corresponds to a planet and reflects the model-computed likelihood of each atmospheric class following interpolation, normalization, and smoothing procedures applied to the observed spectra. Cloudy classifications receive the highest probability across the sample, consistent with the muted or flattened spectral signatures reported in the literature.*

Real-Data Preprocessing

Observed spectra were interpolated onto the same wavelength grid used for synthetic spectra, normalized using the same mean-flux scaling, and smoothed with a Gaussian kernel matched to the synthetic noise distribution.

RESULTS

The random forest classifier achieved approximately 99–100 percent accuracy on the synthetic test set. Most misclassifications occurred between clear and cloudy spectra with heavily flattened features, consistent with overlapping spectral characteristics.

Table 1 presents model predictions for the five observed exoplanets. For each planet, the classifier favored a cloudy classification, with predicted probabilities ranging from 65 to 89 percent. The model assigned substantially lower probabilities to clear and hazy classes. These trends agree broadly with prior studies showing that these planets tend to exhibit muted or featureless spectra due to cloud layers (7-9).

A probability bar chart summarizing the classifier outputs for the five planets is provided (Figure 2), and a confusion matrix (Figure 1) illustrating model performance on the synthetic test set is included to show class-specific accuracy.

DISCUSSION

The results presented in Table 1 show that the random forest classifier was successfully able to identify the five planets it was tested on. The success rate of 100% indicates that the ML model was successfully able to identify the spectral features that the atmospheric classes are differentiated by. The consistent cloudy predictions align with prior observational studies of these exoplanets. For example, the transmission spectra of HD 209458 b, WASP-17b, and WASP-12b have revealed muted absorption features indicative of cloud layers (8). Similarly, the relatively featureless spectra observed for K2-18b and HAT-P-26b are consistent with significant cloud coverage (7, 9). This agreement with the literature supports the idea that the model is capturing real, physically meaningful trends rather than arbitrary patterns.

Limitations must be addressed, however. First, the training data was generated synthetically, meaning the model was not trained on real data. Specifically, only H_2O and CH_4 features were generated, whereas real spectra may include CO, CO_2, NH_3, and other species. Additionally, real spectra contain correlated noise, stellar contamination, and instrument systematics not captured

Table 1. *Machine Learning Model Predictions of Exoplanet Atmospheric Conditions*

Exoplanet	Predicted Label	True Label	Clear	Cloudy	Haze
K2-18b	Cloudy	Cloudy	0.180	0.650	0.170
HD 209458 b	Cloudy	Cloudy	0.250	0.720	0.030
WASP-17b	Cloudy	Cloudy	0.335	0.665	0.000
WASP-12b	Cloudy	Cloudy	0.030	0.885	0.085
HAT-P-26b	Cloudy	Cloudy	0.315	0.650	0.035

in synthetic training data. These domain differences may shift the classifier's decision boundaries.

Second, only five samples were tested, which limited the statistical confidence in the model's generalizability. The small test sample further limits the statistical strength of the conclusions. These targets were intentionally chosen to span a range of planetary types, masses, gravities, and temperatures, allowing the model to be evaluated on diverse cases with minimal data. While this strategy is reasonable for an initial study, the results should be viewed primarily as a proof of concept rather than as a final, statistically validated demonstration of model performance. Substantially larger and more varied datasets will be required to rigorously assess generalizability.

Third, spectral variability among planets, including temperature differences, surface gravity, cloud-top pressures, and metallicity, can also influence the appearance of molecular absorption features and thus classification confidence. Because random forest probability outputs are not well calibrated, the predicted probabilities serve as qualitative indicators rather than rigorous measures of likelihood.

Despite these caveats, the results illustrate the promise of ML for atmospheric classification, as shown through the model's rapid classification ability. Future work could strengthen this approach by incorporating more realistic/real training data, additional molecular species, and a more diverse set of exoplanetary data. These steps together would move the method from a preliminary demonstration toward a more robust and broadly applicable tool for exoplanet atmospheric classification.

This work is a proof-of-concept demonstration and should not be interpreted as a validated atmospheric classifier. The small test sample, simplified molecular assumptions, lack of probability calibration, and synthetic-to-real domain shift all limit the scientific interpretability of the predictions. Larger, more diverse datasets and domain-adaptation techniques will be essential before such classifiers can be used in robust scientific pipelines.

CONCLUSION

This study demonstrates the potential of ML as a useful tool for exoplanet atmospheric classification. Trained on synthetic spectra, the ML model was successful in categorizing real exoplanets with high confidence based on their atmospheric conditions, specifically whether they were clear, or had a presence of clouds and haze. The predictions were in agreement with past findings (7-9), illustrating that the model was able to correctly identify the spectral features of the exoplanets.

This ML model could be used to aid in determining the presence of clouds and haze in exoplanet atmospheres, thereby streamlining the process. Instead of classifying atmosphere types by hand, scientists could instead automate this process by feeding transmission spectrum data into the model. Moreover, the random forest's feature importance rankings may help scientists identify which wavelength regions are most diagnostic for cloud presence.

Future work could adapt the ML model to be able to handle hundreds, or thousands of spectra at a time instead of just one at a time that the preliminary model is able to handle. This would expedite the process even more by being able to run multiple thousands of spectra at one time. By streamlining this process, researchers would be able to allocate more time to other tasks.

The results presented here provide a preliminary but promising demonstration of how ML can complement traditional techniques in exoplanetary science. As observational data continues to expand in quantity, models such as the one developed here will be essential for categorization tasks. As exoplanet datasets continue to expand with JWST and ARIEL, models like this one will be vital for rapid atmospheric classification and for prioritizing targets for detailed retrieval.

DATA CODE AND AVAILABILITY

The code and data are available from the corresponding author upon reasonable request.

FUNDING SOURCES

The author has no funding sources that supported the research and preparation of this article.

CONFLICT OF INTEREST

The author declares that there are no conflicts of interest related to this work

REFERENCES

1. Madhusudhan N. Exoplanetary atmospheres: key insights, challenges, and prospects. *Annu Rev Astron Astrophys*. 2019; 57: 617–63. doi:10.1146/annurev-astro-081817-051846
2. Christiansen JL, McElroy DL, Harbut M, Ciardi DR, Crane M, Good J, *et al*. The NASA Exoplanet Archive and Exoplanet Follow-up Observing Program: data, tools, and usage. *Planet Sci J*. 2025; 6 (8): 186. doi:10.3847/PSJ/ade3c2
3. Rukdee S. Instrumentation prospects for rocky exoplanet atmospheres studies with high resolution spectroscopy. *Sci Rep*. 2024; 14 (1): 27356. doi:10.1038/s41598-024-78071-5
4. Mollière P, Wardenier JP, van Boekel R, Henning T, Molaverdikhani K, Snellen IAG. petitRADTRANS: a Python radiative transfer package for exoplanet characterization and retrieval. *Astron Astrophys*. 2019; 627: A67. doi:10.1051/0004-6361/201935470
5. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al*. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011; 12: 2825–30.
6. Rodriguez DR. The Exo.MAST portal: an exoplanet focused view into MAST data. In: Pizzo R, Deul ER, Mol JD, de Plaa J, Verkouter H, editors. Astronomical Data Analysis Software and Systems XXIX. Vol 527. San Francisco: ASP; 2020. 351.
7. Benneke B, Wong I, Piaulet C, Knutson HA, Lothringer J, Morley CV, *et al*. Water vapor and clouds on the habitable-zone sub-Neptune exoplanet K2-18b. *Astrophys J Lett*. 2019; 887 (1): L14. doi:10.3847/2041-8213/ab59dc
8. Sing DK, Fortney JJ, Nikolov N, Wakeford HR, Kataria T, Evans TM, *et al*. A continuum from clear to cloudy hot-Jupiter exoplanets without primordial water depletion. *Nature*. 2016; 529 (7584): 59–62. doi:10.1038/nature16068
9. Wakeford HR, Wilson TJ, Stevenson KB, Lewis NK. Exoplanet atmosphere forecast: observers should expect spectroscopic transmission features to be muted to 33%. *Res Notes AAS*. 2019; 3 (1): 7. doi:10.3847/2515-5172/aafc63