# Classifying High-Redshift Galaxies from the HETDEX Survey Using a Random Forest Classifier


**Nicholas Davila[1], Oscar Chavez[2], Gene Leung[2], Steven Finkelstein[2]**

[1]Department of Physics, College of Natural Sciences, The University of Texas

[2]Department of Astronomy, College of Natural Sciences, The University of Texas


## Introduction

We trained a machine learning algorithm to autonomously classify large astronomical datasets for high-redshift galaxies. We used raw data from the Hobby-Eberly Telescope Dark Energy Experiment (HETDEX) third internal data release (HDR3) and extracted light spectra to train a Random Forest classification algorithm. We manipulated the spectral data to construct a sample of 10,000 true astrophysical spectra (high-redshift sources) and 10,000 noise sources. The high-redshift data set was constructed by using visually inspected spectra from the GEVIP internal HETDEX detections catalog. The noise sample was constructed by extracting spectra at random positions in the sky where there were no astrophysical objects within 200 arcseconds of the location of extraction. Both samples were then combined into a single dataset and labeled. We labeled a high-redshift source as a '1' and a noise source was labeled a '0'. Using the 'sklearn' package, the entire dataset was split into 70% training and 30% testing sets. The algorithm was trained on the training set, and then employed to classify a source as either a '1' or a '0' in the testing set. The algorithm was able to achieve an accuracy of ~98.55% in its classification predictions with a precision of ~98.19% and a recall of ~98.94%.

## Motivation

Machine learning methods like clustering and classification are very popular in the astronomy machine learning community. However, many ML algorithms struggle with differentiating between noise spectra and true astrophysical (high-redshift in this case) spectra. The motivation for this project was to implement an algorithm that focused on solving the astrophysical source vs. noise problem specifically. This will allow for more high-redshift sources to be studied, which will help us learn more about the period of reionization in the universal timeline.

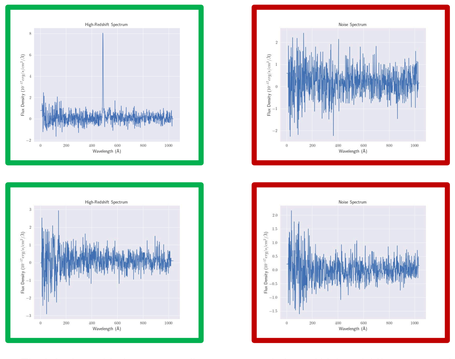



*Figure 1: The left plots with the green outline are spectral plots of high-redshift true astrophysical objects. The green is to highlight that those are the spectral plots we want. The right plots with red outlines are spectral plots of noise. The red outline is to highlight that (ultimately) we do not want those, since studying noise is not interesting to us right now.*

## Methods and Results

**Data Selection**

1. Extracted 10,000 'high-confidence' high-redshift spectral data from a GEVIP HETDEX detection catalog based on signal-to-noise ratio values greater than 7 and a pLya score larger than 0.95
2. Extracted 10,000 noise sources spectral data at random locations in the sky after checking that there we no sources within 200 arcseconds of the extraction point

**Data Manipulation**

1. Removed noisy areas from high-redshift sources by dividing the sources by their respective spectral error
2. Combined the noise and high-redshift data sets into one large 20,000 sized data set and labeled high-redshift sources as '1' and noise sources as '0.'
3. Using sklearn train_test_split the data was split into 70% training and 30% testing sets

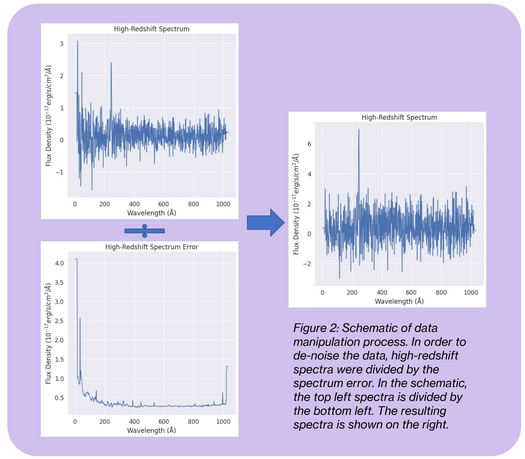
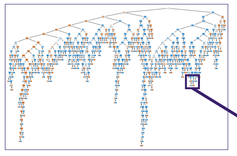



*Figure 2: Schematic of data manipulation process. In order to de-noise the data, high-redshift spectra were divided by the spectrum error. In the schematic, the top left spectra is divided by the bottom left. The resulting spectra is shown on the right.*



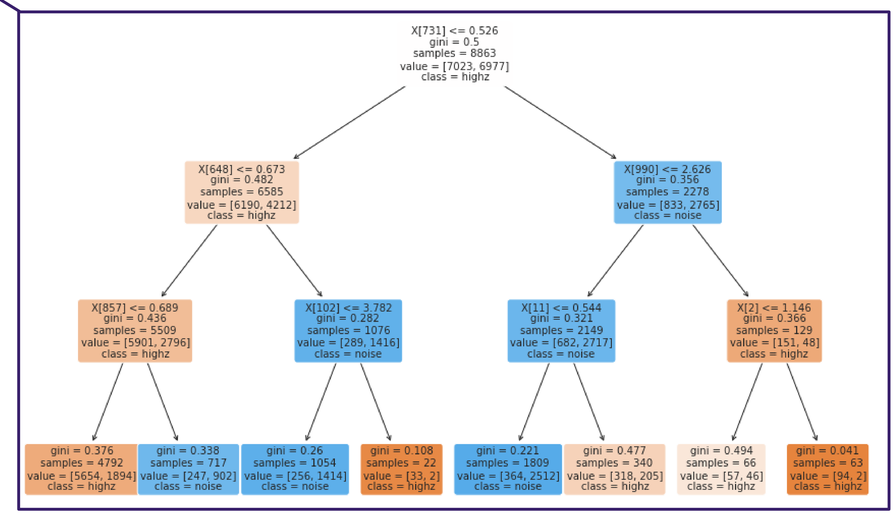



*Figure 3: On the top left is a zoomed-out plot of the first decision tree created by our random forest classification algorithm (out of 700 total). The bottom right plot is a zoomed-in plot into the tree where the information on why splits were made is visible. The first line in all the boxes is the feature name with the split value. The split value is decided by the algorithm after selecting a threshold value which gives the highest information gain for that split. The second line 'gini' is the deciding factor to select the best split value or select the feature at the next node etc. The third line 'samples' is the remaining number of samples at that particular node. The fourth line 'values' is the number of each class remaining at that particular node.*

**Strengths:**

- RF is an ensemble learning method which combines multiple different techniques (sometimes per each node) to improve accuracy
- Very robust to outliers and overfitting
- Easy to implement and popular so there is a lot of documentation for help

**Weaknesses:**

- Computationally expensive and slow when more trees and data are involved
- Very dependent on the amount of data had. Clear increasing trend between higher accuracy and larger amount of data
- Difficult to interpret hundreds of trees/decisions

## Results

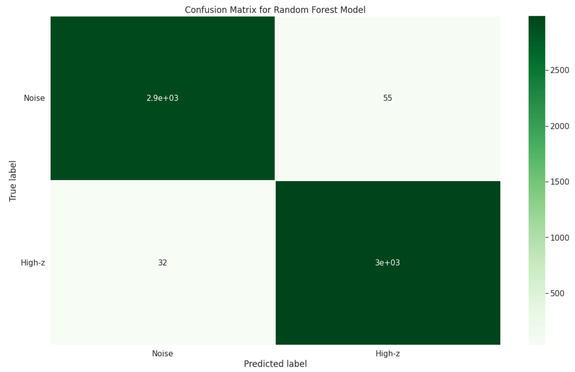



*Figure 4: Confusion matrix for the Random Forest model. The top left square tells us our true negatives (predicted noise classified correctly as noise). The bottom right square tell us our true positives (predicted high-redshift classified correctly as high redshift). The top right square tells us our false positives (predicted high-redshift but is noise), and the bottom left square tells us our false negatives (predicted noise but is high-redshift).*

We split the combined 20,000 sized data into a 70% training and 30% testing set. Out of the 6,000 sources that were in the testing set, 2985 were correctly identified as high-redshift, 2928 were correctly identified as noise, 55 were high-redshift sources incorrectly identified as noise and 32 were noise sources incorrectly identifies as high-redshift.

## Conclusion

We trained a Random Forest machine learning algorithm to distinguish between true astrophysical spectra and noise spectra. First, we extracted raw data and reduced the noise of our target data by dividing spectral error data out of the true astrophysical data. Applying the random forest classifier on the denoised data allowed us to achieve a classification accuracy of ~98.55%. In the future, this algorithm could be used to support algorithms which usually suffer with high-redshift vs. noise classification.

## References


Mentuch Cooper, E. et al, "HETDEX Public Source Catalog 1: 280K Sources including over 50K Lyman Alpha Emitters From an Untargeted Wide-area Spectroscopic Survey"


## Acknowledgments


Thank you to Oscar Chavez, Dr. Steven Finkelstein, Dr. Gene Leung, and GEVIP for their support throughout my research. The Vertically Integrated Project (VIP): Galaxy Evolution group gratefully acknowledges the support from National Science Foundation (NSF) grant AST-1908817.