Using Machine Learning to Classify Lyman-α-Emitting Galaxies in the NEP Field

Nicholas Davila[1], Valentina Tardugno Poleo[1], Delaney White[2], Oscar A. Chavez Ortiz[2], Gene K. Leung[2], Steven F. 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

The University of Texas at Austin
Department of Astronomy
College of Natural Sciences

HETDEX Hobby-Eberly Telescope Dark Energy Experiment Illuminating the Darkness

GEVIP

Abstract

One of GEVIP's (Galaxy Evolution Vertically Integrated Project) themes is working with HETDEX (Hobby-Eberly Telescope Dark Energy Experiment). HETDEX is an unbiased spectroscopic survey using the 10m Hobby Eberly Telescope (HET) and its VIRUS integral-field unit (IFU) spectrograph. HETDEX is in the process of discovering distant galaxies on the basis of their strong Lyman-α emissions. In some GEVIP projects, we use the discovered Lyman-α emitting galaxies with the goal of understanding how the Milky Way galaxy was formed. In order to get usable data, we work on classifying Lyman-α emitting galaxies from large sets of data which contain different astronomical objects. To classify, we divide astronomical objects into groups based on their visual appearance. However, data in astronomy is getting larger and more complex, so we are turning to machine learning algorithms that can adapt to increasingly large sets of data. Therefore, this project aims to create a machine learning algorithm to accelerate the classification of Lyman-α galaxies. Our first step is to explore different machine learning algorithms related to classification. We then want to compare the algorithms and find which one is best suited for our purposes. We want an algorithm that takes images as input and can output a Lyman-α score between 0 and 1 (a probability the source is Lyman-α, where 0 is not probable at all and 1 is very probable). Lastly, we want to implement our chosen algorithm to the smaller NEP field data, and then optimize it, to one day use it with other data sets.

Research Goal

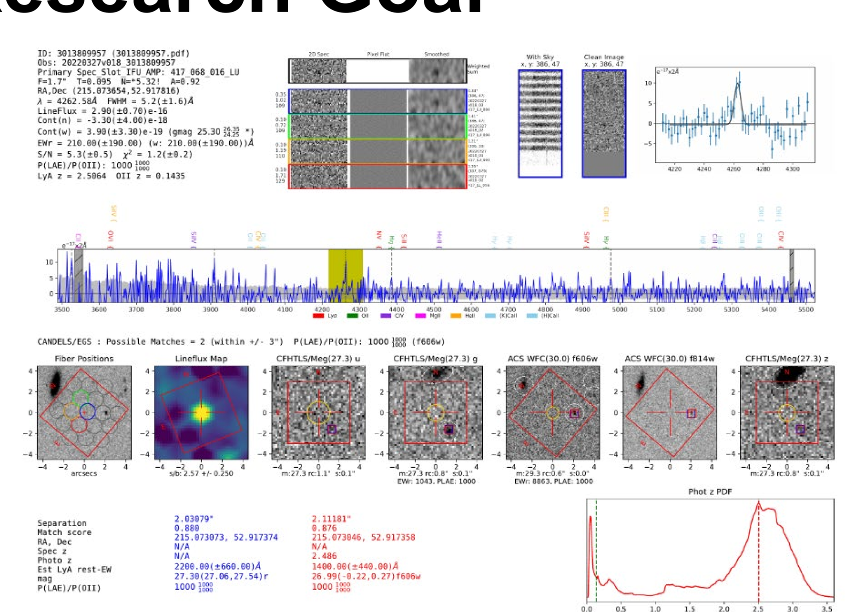

Figure 1: Elixer plot used to manually classify astronomical objects.

- Our goal is to develop a deep learning algorithm that can help us move away from manually classifying galaxies with elixer plots.

Methods and Expected Results

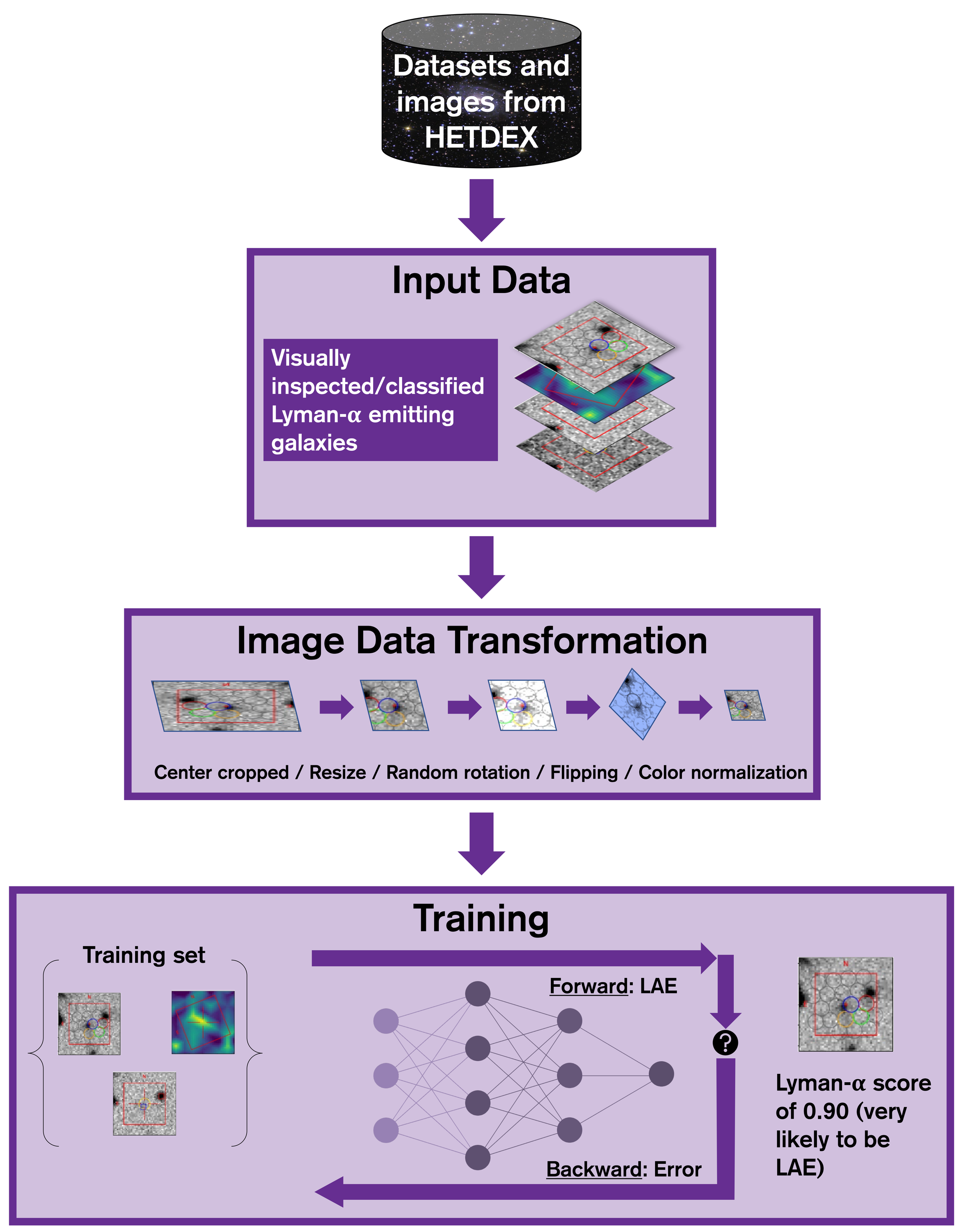

***Figure 2:** Training architecture for deep learning galaxy classification. We input sources that we are confident are LAEs and we also input some random sources. Then we do some type of data transformation - some examples are listed in the figure. Then we train the algorithm until we get the highest possible score (score does not improve). If the score is not over 0.95 on all known sources, then we reconfigure.*

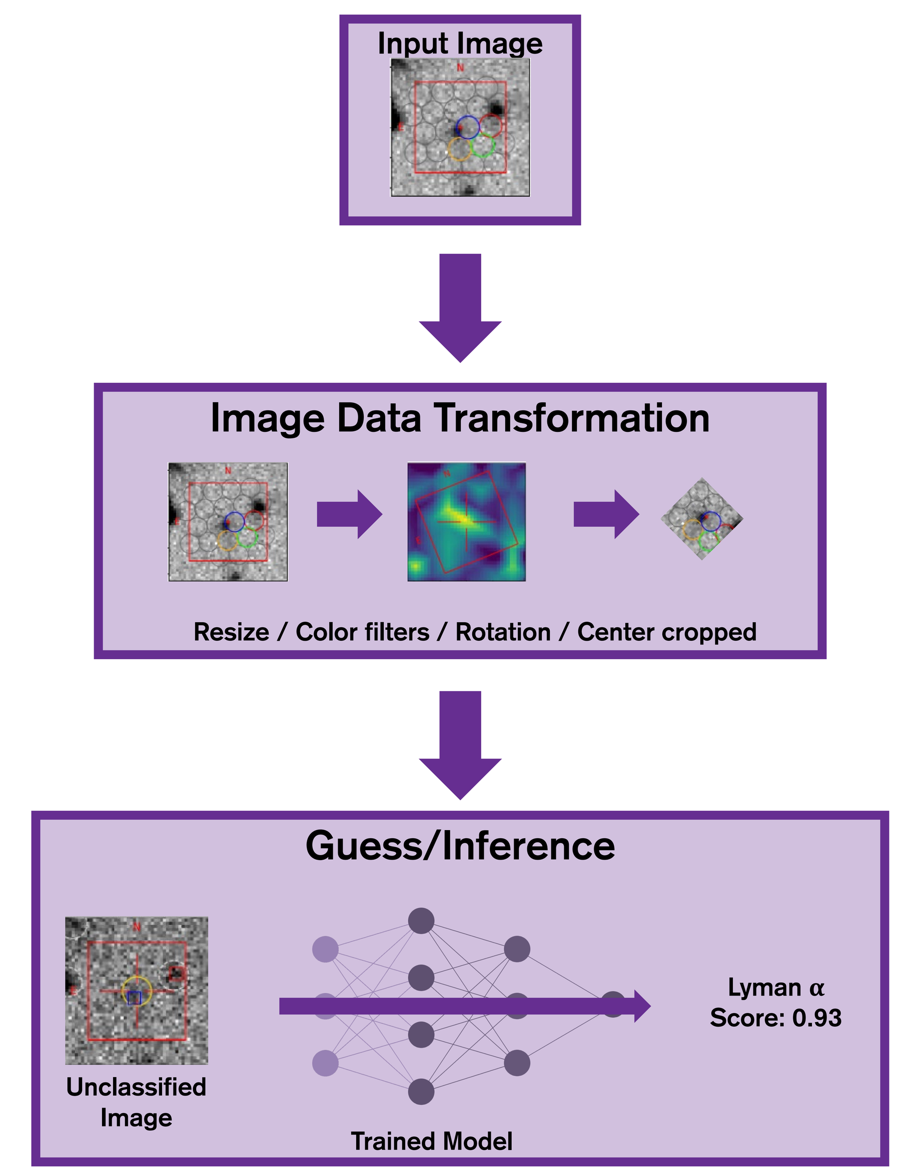

***Figure 3:** Use case architecture. This is the framework for how the algorithm should work once trained. We give it an unclassified (by humans) image or images. The algorithm manipulates the image data in some way (to make classifications faster and more reliable). Then the algorithm uses its experience i.e., training to classify the image and give it a Lyman α score. The score is a percentage value which indicates the likelihood that the astronomical object is a LAE. The closer the score is to 1, the more likely it is that the object is a LAE. For example, a score of 0.93 indicates the object analyzed by the deep learning model is a LAE.*

Right now, the number one choice is to train a convolutional neural network (CNN) so that it can distinguish between real LAEs and noise images. One of the main advantages of a CNN is that it can automatically detect important features without any human supervision. So, our strategy is to select sources as LAE candidates that have been classified as real LAEs by the GEVIP group and select sources that have been classified as not LAE.

Training:

- For the training of the CNN, we will create data sets of LAEs and non-LAEs. One of the main problems of creating a classification algorithm is the number of correctly identified LAEs is relatively limited. To mitigate this, we can simulate images of LAEs (by using the SynPipe software (Yoshiaki et al. 2021)) or classify more objects by hand.
- Once trained, we want to give the algorithm a data set with objects we are positive are LAEs. We want to let the algorithm train until its accuracy cannot go any higher. We are going to grade its accuracy by how high the Lyman-α scores are when it classifies an object we know to be a LAE. The score will be a percentage value which indicates the likelihood that the astronomical object is a LAE. The closer the score is to 1, the more likely it is that the object is a LAE. For example, a score of 0.93 indicates the object analyzed by the deep learning model is a LAE.

Methods Cont.

Use case:

- We want the algorithm to be simple to use. Input will be an image or a set of images. The algorithm transforms the image data for the best optimization. The algorithm then classifies the sources and returns all the sources Lyman-α scores and even a list of which ones are above a given range.

Future Work

- We hope to develop a machine earning algorithm to classify LAEs from given data sets of astronomical objects. The algorithm would ideally classify LAEs with an accuracy of over 90%.
- In the future we hope to develop the algorithm and use it to get more usable data (quicker) for all the GEVIP projects that require LAEs.

Acknowledgements

Thank you to Dr. Steven Finkelstein, Dr. Gene Leung and Oscar Chavez for supervising and supporting the project. Also thank you to Delaney White and Valentina Tardugno Poleo for their feedback throughout the entire research process.

References

Ono, Y., Itoh, R., Shibuya, T., Ouchi, M., et al. (2021). SILVERRUSH X: Machine Learning-aided Selection of 9318 LAEs at z= 2.2, 3.3, 4.9, 5.7, 6.6, and 7.0 from the HSC SSP and CHORUS Survey Data. The Astrophysical Journal, 911(2), 78.