Spurious Source Rejection Algorithms for Galaxies from JWST

Marissa Perry[1], Lipika Chatur[1], and Urvi Thakurdesai[1]
The University of Texas at Austin

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

INTRODUCTION

Studying the reionization era is valuable to gain insight into the assembly of the early universe and its evolution to present time. However, for such analyses big data has transformed the feasibility of quality control during sample selection.

In this work, two types of spurious source rejection algorithms were applied to samples of high-redshift (z = 3 - 17) galaxies from JWST (James Webb Space Telescope) NIRCam photometry.

A spurious source is a false detection of a galaxy. One common type that contaminates samples is a diffraction spike (stellar artifact).

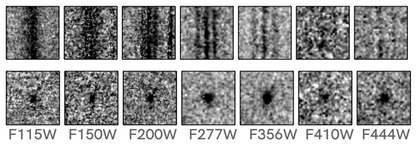


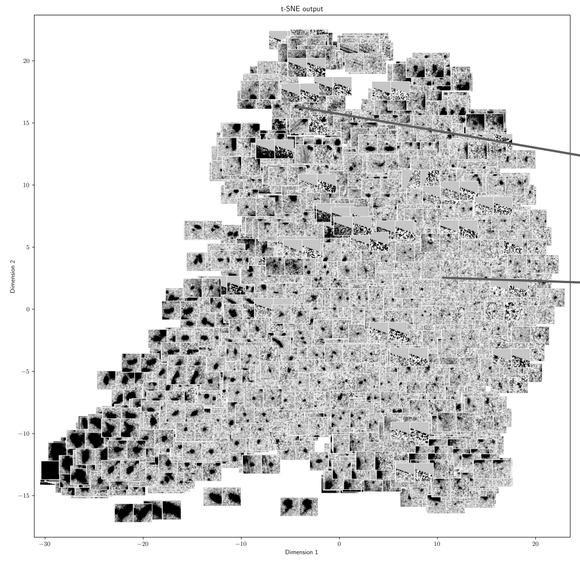
Fig. 1. Top: diffraction spike spurious source. Bottom: real detection of a galaxy source

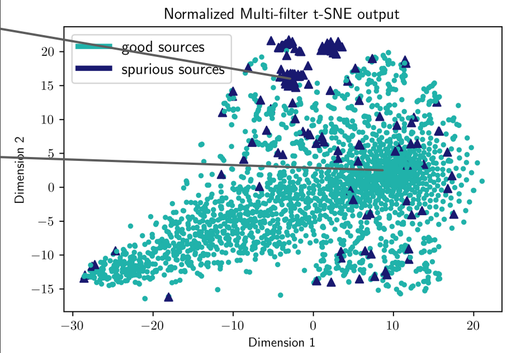
ML CLUSTERING ALGORITHM

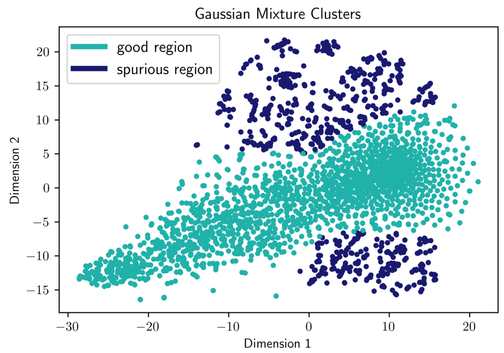
A sample of 2,310 sources from the NGDEEP catalog was visually inspected and assigned a "real" or "spurious" label. This sample contained about 7.7% spurious sources (Fig.3).

We took a preprocessing step to normalize the data by diving out by the maximum value. Then, we applied a data mining task on our dataset, for which we used the t-Distributed Stochastic Neighbor Embedding (t-SNE) dimensionality reduction technique (Fig. 2 and Fig. 3).

Lastly, we implemented a clustering task to our dimensionally reduced dataset. For this, we used the Gaussian Mixture Model (GMM) clustering algorithm. Although this ML algorithm is unsupervised, we were able to choose a reasonable number of clusters for the algorithm to split up our dataset. We found that 5 clusters yielded the best clustering of spurious sources (Fig. 4).







In Fig. 4, the two dark blue clusters represent two spurious regions containing a total of 695 sources, out of which 140 spurious sources lie within.

Cluster 1: 217 sources, 6% spurious sources.
Cluster 2: 478 sources, 27% spurious sources.

Fig. 2. Left: t-SNE dimensionally reduced 2,310 NGDEEP sources. Each source's position is determined from data in all filters shown in Fig.1, and are represented visually by images of the F150W and F277W filters. Fig. 3. Middle: labeling of t-SNE dimensionally reduced sources . The dark blue triangles represnt the "bad" sources. Fig. 4. Right: GMM clustering model implemented with 5 clusters. 2 clusters are grouped together in this figure to represent a significantly spurious region. These clusters were chosen for exhibiting >5% spurious sources. The dark blue regions make up 6% of the total dataset.

DEEP LEARNING CNN ALGORITHM

With the same sample from NGDEEP, Keras (the high-level API of TensorFlow) was used to train a convolutional neural network (CNN) binary classifier. This algorithm focused on identifying diffraction spike and chip edge sources, which comprised of 4.81% of the sample (111 sources).

To combat this class imbalance, the model was given higher weight for correctly identifying spurious images. Additionally, the photometric images were normalized to scale the pixel values between zero and one, allowing the neural network to converge quicker.

flatten + dense + dropout + dense + output neuron

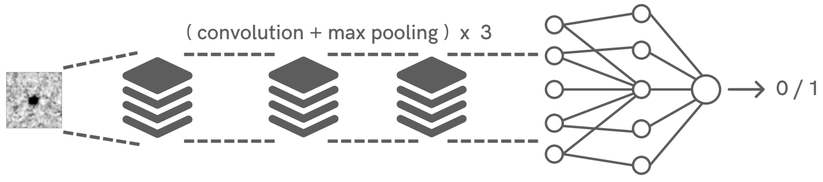


Fig. 5. CNN architecture. All layers utilize a rectified linear unit (ReLU) activation function, except for the output neuron operating with a sigmoid activation function.

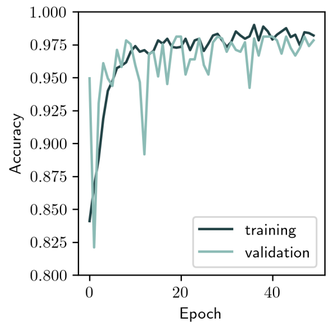


Fig. 6. Performance of the training process. The model was then presented with unseen images (a validation set) and evaluated once more. It performed with an accuracy of 97.84% on the validation set.

The trained CNN model was then tested on unseen and unlabeled data, using a sample of 13,975 galaxy sources from CEERS. A cut was made where predictions less than 70% confident in the source being real were classified as spurious. The algorithm found the sample to contain 1.02 % spurious sources (142 sources).

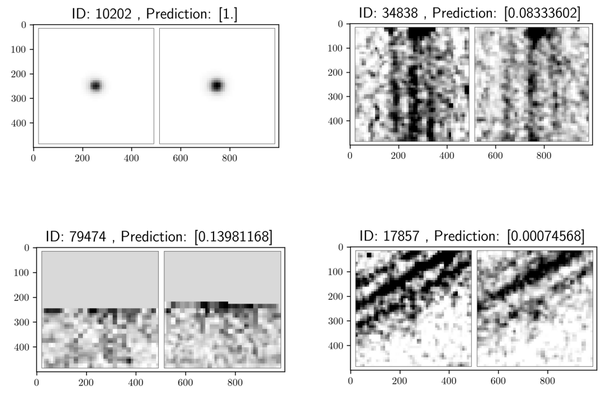


Fig.7. Example predictions on CEERS dataset. top left: good source, bottom left: chip edge, top right: diffraction spike, bottom right: diffraction spike

RESULTS

Comparing both algorithms, the ML clustering algorithm is an appropriate catching net for spurious sources in a given sample. However, implementing this kind of algorithm will only reduced the volume of visual inspection during a sample selection process. Whereas the CNN algorithm, if trained well enough, potentially removes the need for visual inspection altogether.

Within the next decade, ML and deep learning will become critical to deal with the volume of data expected for upcoming technologically advanced Earth and space telescopes, such as the Giant Magellan Telescope (GMT) and Vera Rubin Observatory. In many cases, ML and deep learning data visualization tools can help boost productivity and thus discoveries of the unknown.

ACKNOWLEDGEMENTS

We would like to thank Dr. Steven Finkelstein for providing the JWST NGDEEP and CEERS dataset. The Vertically Integrated Project (VIP) Galaxy Evolution group gratefully acknowledges the support from National Science Foundation (NSF).