UPDATED MAY 2, 2021

Systematically Searching for VLITE Transients

UJWAL KUMAR[1]

[1]*The George Washington University*
Washington DC, USA

ABSTRACT

The detection of low-frequency radio transients can help pinpoint high-energy events in our sky such as supernovae or gamma-ray bursts. We use data from the VLA Low Band Ionospheric and Transient Experiment (VLITE) to search for such transients. By running VLITE images through the LOFAR Transients Pipeline (TraP), an algorithm that measures the fluxes and positions of sources that it finds in images, and quantifies each sources' variation in flux over time into two variability parameters, η and V, we can begin to search for such transients. To do this, we employ two machine learning strategies, anomaly detection and logistic regression, to classify these sources into stable sources and transient candidates. We find the logistic regression strategy to yield more promising results for VLITE data and we also provide an estimate of $\approx 40\%$ systematic flux uncertainty for future VLITE transient searches.

Keywords: Radio Astronomy – Machine Learning – Transients

1. *Introduction*

Detecting sources in our sky that display large changes in brightness over timescales from seconds to years – sources referred to as *transient* sources – can pinpoint towards high-energy events such as supernovae in the universe at various distances or redshifts. These events can be detected across the electromagnetic spectrum and can provide information about the the behaviour of physics in extreme conditions that cannot be recreated on Earth.

The different regimes of the electromagnetic spectrum reveal varying properties about different objects in the universe and the spaces in between. For example, infalling gas and dust onto compact objects (neutron stars or black holes) can release X-ray emission and this emission can reveal properties about the media surrounding the compact object or about the compact object itself. In this paper, we aim to study and find transients in the low frequency radio spectrum. Looking in the radio allows us to probe faster moving ejecta such as the jets from active galactic nuclei (AGNs) and can pinpoint towards other such high energy events like gamma ray bursts (GRBs). LOFAR in the Netherlands and the Jansky Very Large Array Low-band Ionosphere and Transient Experiment (VLITE) on the Very Large Array (VLA) in the United States are examples of radio telescope interferometers used to conduct low-frequency radio surveys (van Haarlem et al. 2013). Interferometry is a special technique used in radio imaging that improves the resolution of the image. According to the Rayleigh criterion, two sources can be resolved only if they are separated by a certain angular distance that is proportional to the wavelength and inversely proportional to the diameter of the radio dish. Radio wavelengths therefore require larger dishes to compensate for the long wavelengths. As building a large dish is impractical for multiple reasons, interferometry uses multiple telescopes in an array that scan the sky in such a way that the effec-

tive diameter of the dish is the distance between the farthest two telescopes. The VLITE instrument, developed by the Naval Research Laboratory (NRL) and the National Radio Astronomy Observatory (NRAO) builds on VLA interferometry capabilities to essentially create a two-in-one telescope. The VLITE system operates independently of the VLA system with a large field of view of 5.5 deg^2 allowing for greater probabilities of detecting transients.

In this work, we utilize VLITE data processed through the LOFAR Transients Pipeline (TraP). TraP is a tool developed to study LOFAR data but can be deployed to other surveys. It takes in images as inputs and processes them through a series of steps – it searches for sources in each image, measures their fluxes and positions, and builds light curves that measure the flux over time. It then outputs two variability parameters, η_ν and V_ν, that quantify the changes in brightness. These parameters and the entire pipeline work are presented in Swinbank et al. (2015).

Given a list of sources with their variability quantified, our next goal is to classify these sources into stable and transient sources. Machine learning techniques have long been developed to classify data and in order to do so, these techniques use existing or training data to make predictions and/or classifications. These techniques or algorithms can be divided into supervised and unsupervised machine learning algorithms, the former of which uses input labels to classify data and the latter of which gives no input labels and allows the computer to decide what the labels or categories are. In this work, we aim to employ two supervised machine learning strategies, anomaly detection and logistic regression, developed and presented in Rowlinson et al. (2019) to distinguish potential transients from stable sources. The training data we use is VLITE data injected with simulated transients and these data were created by Alexandra Weikert, a previous student at The George Washington University, in collaboration with the NRL. These strategies will use the variability parameters outputted by TraP as well as other variables such as the maximum flux to classify these sources and present transient candidates. The goal of creating such data is that we can run machine learning algorithms against pre-labelled VLITE data where we know the transients from the stable sources. This thereby allows us to test the accuracy and efficacy of the machine learning strategies.

Lastly, we aim to estimate levels of systematic flux uncertainties in VLITE transient searches. Systematic uncertainties are errors in measurements that stay constant from measurement to measurement and arise due to constraining effects of the tools used or in our case, arise largely from calibration errors in the telescope or instruments. When accounted for, however, estimates of systematic uncertainties can be included in the analysis of data for a more accurate understanding of where our data truly lie. Rowlinson et. al describes the TraP process and notes that the flux measurements of sources do not include VLITE imaging systematic uncertainties. This provides our motivation for understanding the levels of systematic flux uncertainties in the data.

Through this paper, we aim to build on work presented in Rowlinson et. al by estimating the systematic uncertainties in flux measurements made by TraP, and aim to apply machine learning strategies to identify potential transients in VLITE data. This paper will first present the outline of TraP and the system through which the images are processed. We then present an overview of the motivation to estimating systematic uncertainties in this data and then outline the backend structure of the anomaly detection and logistic regression strategies. Finally, we present the results of our studies of these systematic uncertainties and of the machine learning strategies. We discuss reasons for variances in results between the two strategies and outline future work to be conducted to gain a better

understanding of how we can use such techniques to search for transient sources in the radio sky.

2. *VLITE Data Processing through TraP*

Images first taken by the VLITE instrument on VLA are run through the TraP algorithm initially developed for LOFAR data. Figure 1 outlines the TraP process showing the inputs, the pipeline processing, and the data products. When images are inputted into TraP, a sourcefinder package is deployed to detect sources of light in each image. For each source detection, it measures the flux and the position of the source with their respective uncertainties. This process is repeated for all the images and builds a light curve measuring the flux over time. Finally, TraP quantifies the changes in variability for each source into two variability parameters, η and V.

It should be noted that the variability of the source can be analyzed at different timescales. A source may display sharp changes in brightness over short timescales but these changes may be insignificant when analyzed over the timescales of days to months. However, sources that are variable at different timescales often hint at the type of source we are looking at (Pietka et al. 2014) In this work, light curves can be sliced into diurnal intervals or 10-minute intervals.

Once run through TraP, a repository named Banana is used as a web interface to display the source ID, the fluxes and positions, and the two variability parameters. We scrape Banana to then study the graphical representations of the variability parameters (an example of which is seen in Figure 2) and other variables such as the max/median flux.

3. *Estimating Levels of Systematic Uncertainties*

Systematic uncertainties arise in all types of data involving measurements. These are uncertainties that arise largely from errors in the calibration of instruments and tools used or may arise from other errors that stay the constant from measurement to measurement, i.e., they are not random. Rowlinson et al. (2019) outlines the TraP process but notes that measurements of flux by TraP do not account for VLITE systematic uncertainties and this provides our motivation for estimating systematic uncertainties in our data.

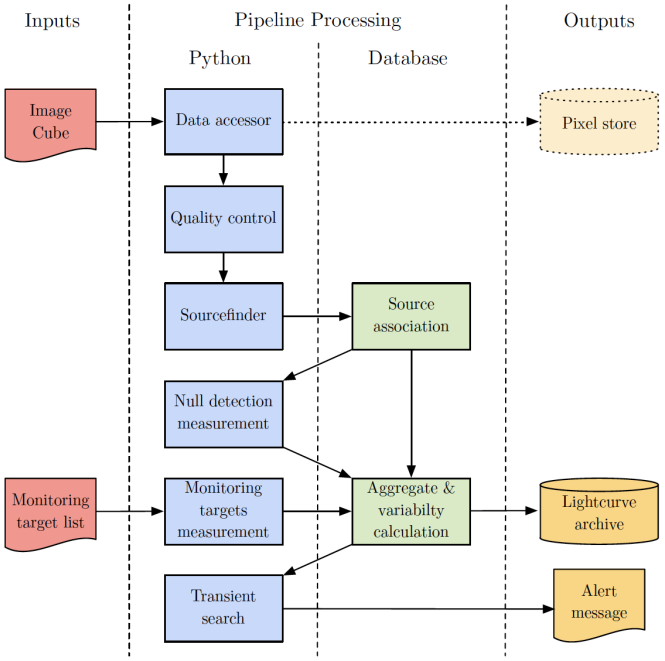

Figure 1: Overview of Transients Pipeline (Swinbank et al. 2015)

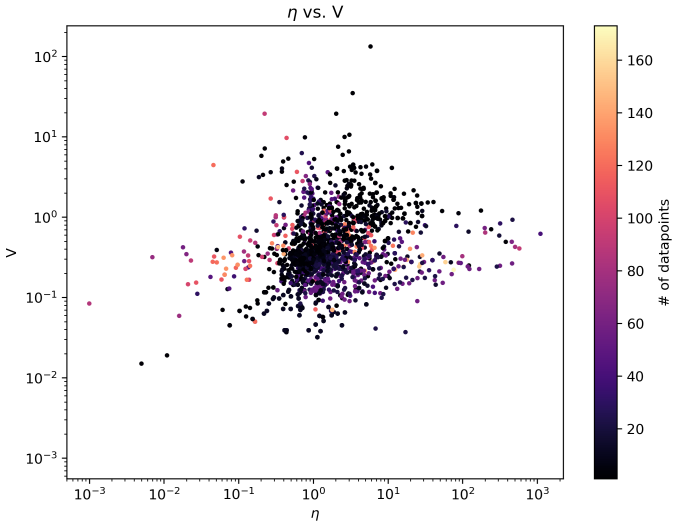

Figure 2: V vs η plotted for VLITE (10-minute interval) data colour mapped by the number of data points in the lightcurve

To estimate the systematic uncertainties, we first look at the formulae for η and V where N indicates the number of measurements or data points in the light curve, $\overline{I_v}^2$ is the mean average of the integrated flux whole squared and $\overline{I_v^2}$ is the mean average of the squares of the integrated flux. The weight on the integrated flux, $\overline{\omega_\nu}$, is defined in Rownlinson et. al as $\frac{1}{N}\Sigma_i \frac{1}{\sigma_i^2}$ where σ is the error on the flux measurement.

$$\eta = \frac{N}{N-1}\left(\overline{\omega_\nu I_\nu^2} - \frac{\overline{\omega_\nu I_\nu}^2}{\overline{\omega_\nu}}\right) \tag{1}$$

$$V_\nu = \frac{1}{\overline{I_v}}\sqrt{\frac{N}{N-1}(\overline{I_v^2} - \overline{I_v}^2)} \tag{2}$$

Given measurements of I_v for each source over time, we can recalculate η by adjusting the weight on the flux such that $\sigma = \sqrt{\sigma_{stat}^2 + \sigma_{sys}^2}$ where σ_{sys} is equal to a fraction or percentage of the integrated flux.

Given that the peak flux should be theoretically invariant when accounting for systematic uncertainties; η, the only parameter that depends on the error, should stay independent of the peak flux. The next two subsections will explore the two machine learning strategies that will be used to identify potential transient candidates and we will discuss in our results the estimates of the systematic uncertainty in these data and the findings of the two machine learning strategies.

4. *Machine Learning Strategies*

4.1. *Anomaly Detection Strategy*

In order to find the transients in the data, we first choose to employ a supervised machine learning strategy, anomaly detection. Anomaly detection attempts to divide the dataset into "normal" and "unusual" sources and in order to do this, the algorithm is trained using the VLITE simulated image training data. This serves as a benchmark for the algorithm to understand what kinds of data are normal and what kinds are unusual.

The backend of anomaly detection measures the Gaussianity of the distributions of η and V and finds data that deviates from these Gaussian distributions to be anomalies. To classify the training dataset, the anomaly detection code aims to find thresholds in η and V such that a datapoint with a value greater than the threshold of both η and V would be transients. That is, we split the plot seen in Figure 2 into four quadrants, the top right quadrant of which must enclose the found transients. Equation 3 takes into account the probability distribution function of η and V and if Π is greater than or less than some chosen value, it is either classified as a transient candidate or a stable source.

$$\Pi = \text{pdf}(\eta)\text{pdf}(V) \tag{3}$$

In order for the algorithm to understand its performance, it measures two statistical parameters after each classification run – the precision and recall – which are defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

Where TP stands for True Positives and refers to the number of correctly identified "anomalous" sources and FP stands for False Positives and refers to the number of wrongly identified "anomalous" sources. FN stands for False Negatives which counts the number of anomalous sources not identified at all. Thus, the precision quantifies the total number of correctly identified anomalies and recall measures the number of anomalous sources identified as a fraction of all the anomalous sources that exist in the data.

The data trains by classifying the training dataset, with a target precision and recall value of 0.95 each, and continues this until the errors in the classification reach a constant value.

We aim to run this anomaly detection code against VLITE images injected with simulated transients (that allow us to cross reference true transients). Once tested, we aim to apply its de-

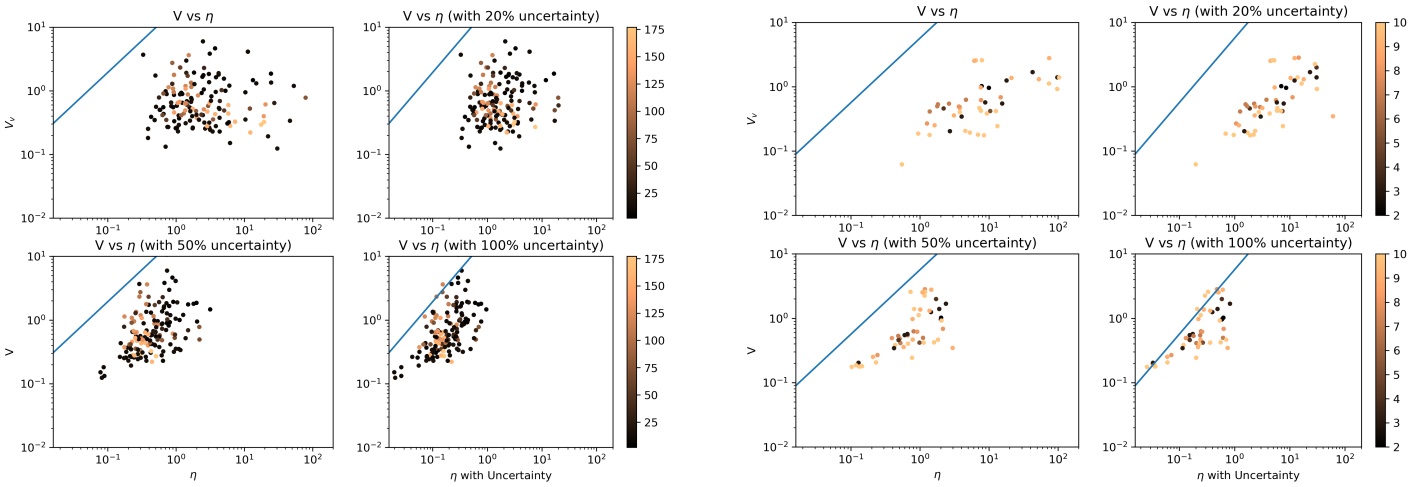

Figure 3: η vs V extracted from VLITE data run through TraP plotted with varying levels of systematic uncertainty in the integrated flux. Colour mapped by number of data points in the light curve.

tection strategies to VLITE data to find transient candidates.

4.2. *Logistic Regression Strategy*

Higher number of input parameters in a model can help machine learning strategies learn more about the data. In the case of anomaly detection, we calculated η and V with the number of images, the integrated flux and the error on the flux. A drawback to the anomaly detection strategy is that it does not consider parameters such as max or median flux. Thus, we employ a logistic regression strategy, which allows for a larger number of input parameters to be accounted for.

The logistic regression algorithm we employ, from Rowlinson et al. (2019), considers four parameters: η, V, the maximum flux, $F_{\rm max}$ and the ratio of the max to the average mean flux, R = $f_{\rm max}/f_{\rm mean}$. We then consider a matrix $\mathbf{X}$ with each column representing the value of the aforementioned parameters for N rows or sources. We also consider a column vector θ such that $\theta \cdot \mathbf{X}$ outputs a scalar value (for each source) that classifies the source as a transient or a stable source. In our work, the logistic regression classifier, seen in equation 6, outputs approximately 0 (stable) or 1 (variable). That is, if $\Sigma = 1$, then the source is a potential transient candidate and if $\Sigma = 0$, then the source is classified as stable. In machine learning, this equation is referred to as the sigmoid function. Visually, the sigmoid function plots a line in a multidimensional space separating potential transients from stable sources with each dimension representing a variable in $\mathbf{X}$.

$$\Sigma = (1 + e^{-\theta \cdot \mathbf{X}})^{-1} \tag{6}$$

The error function which plots the correctness of the algorithm's classifications (based on training data), and can be plotted for larger and larger training data sets till we see no extra information being learnt, i.e., no extra precision is gained through extra computational load. This concept spans across the entire spectrum of machine learning and is extremely useful when classifying multi-parameter space data while also considering computational costs and efficiencies.

From here, we will discuss the findings of systematic uncertainties in data and the results these machine learning strategies provided in our effort to search for transients.

5. *Estimates of Systematic Uncertainty*

While there are numerous methods to estimating systematic uncertainties, our work studies the plots of η against the peak flux while altering the uncertainty, σ (and consequently ω in equation 1). Given the statistical uncertainty on the integrated flux, we

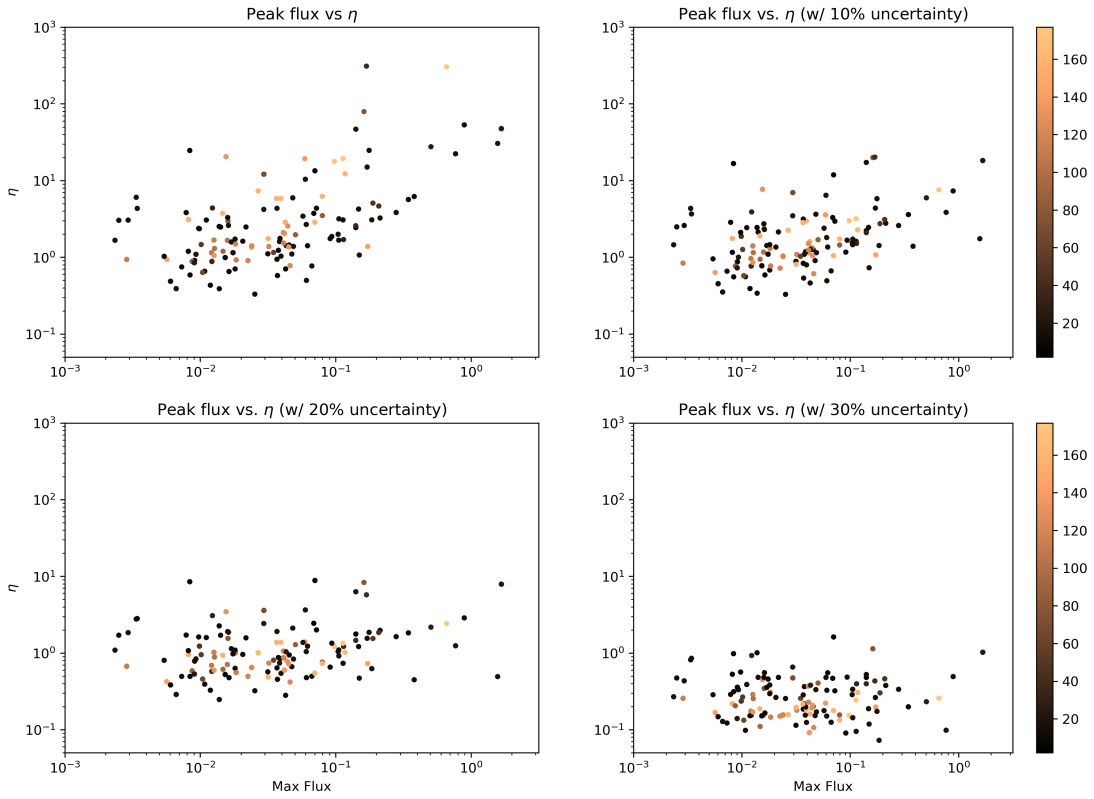

Figure 4: η vs Maximum Flux for η with 0%, 10%, 20%, and 30% uncertainties.

can recalculate η with varying levels of systematic uncertainty.

In order to check the accuracy of our results, we also plot a limiting case combining equations 1 and 2 and setting $\omega_v = 1$. If the uncertainty in the data were equal to the size of the data, we would expect that the data would fall along this boundary and we use this to test our results. This limiting case is given by the equation:

$$V_\nu = \frac{1}{\overline{I_v}}\sqrt{\eta} \tag{7}$$

Figure 3 shows the plots of V vs η for variabilities sliced into all-day intervals (left) and 10-minute intervals (right). In each of these cases, we see that the data lies along the limiting case line indicating that the uncertainties behave as we expect. Given this check on accuracy, Figure 4 plots η against the peak flux with 0-30% uncertainty with step sizes of 10%. Visually, the plots are seen to get flatter in the vertical direction and we also notice that η displays flattening like behaviour as uncertainties reach at least 20%.

To gain a concrete estimate of the flux uncertainty, we can fit a Gaussian to the distribution of η in Figure 4 using Python's `SciPy` package. We then study the spread of η, proxied by the σ value of the Gaussian fit, when plotted against the peak flux.

We plot the σ values of the Gaussian fits to η_v evaluated with varying percentages of systematic uncertainties and see these results in Figure 5. This figure shows that the width is minimized at approximately 40% with higher values percentages resulting in overfitting the data.

6. *Machine Learning Results*

We run the machine learning strategies from sections 4.1 and 4.2 with a target precision and recall of 0.95 as aforementioned against the VLITE simulated data and compare it to results of these strategies run against LOFAR data presented in Rowlinson et al. (2019). In this section, we present the

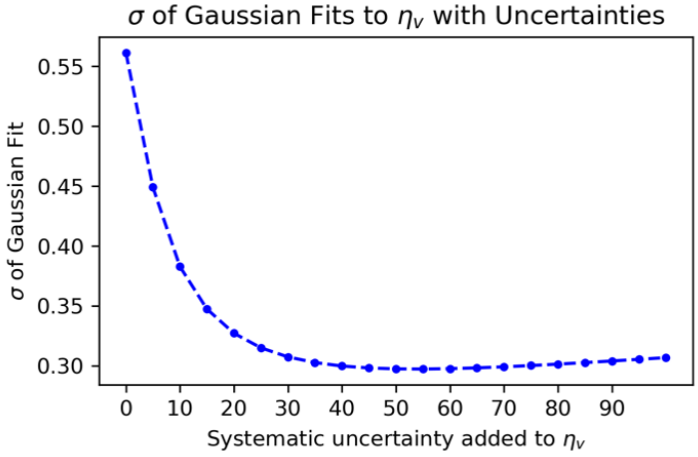

Figure 5: σ value of Gaussian fit to η when plotted against peak flux for varying levels of systematic uncertainty

results of these strategies and discuss reasons for differences in results.

Figure 6 presents the results of the anomaly detection strategy for VLITE (left) and LOFAR (right) data and plots the thresholds of η and V seen with the dashed lines. It plots the correctly identified transients (TP in blue), transients not labelled as transients (FN in green), stable sources incorrectly identified as transients (FP in red) and sources correctly identified as transients seen in the shaded region. Histogram distributions for each parameter are plotted to show the distribution of the variability parameters. Through this strategy, we find a precision of 0.84 and a recall of 0.14 when run against VLITE data. This compares to a precision and recall of 0.90 and 0.86 respectively when run against LOFAR data clearly indicating that the scores for VLITE were low.

Figure 7 shows the results of the logistic regression strategy. We find a precision and recall of 0.93 and 0.68 when run against VLITE data, a marked improvement from the anomaly detection results. This compares to the precision and recall when run against LOFAR data which yields a precision and recall of 0.98 and 0.86 respectively.

Thus, we find that the logistic regression algorithm yielded better values of precision and recall for VLITE data with injected simulated transients (0.93 and 0.68 compared to 0.84 and 0.14 for the anomaly detection strategy). We review reasons for the dramatic differences between the results from the anomaly detection run against LOFAR versus the VLITE data and find that the shape of the data plays a key factor in the results. In Figure 6, we see that VLITE data forks into two "legs" – with high V and median η, and high η and median V. The LOFAR data, on the other hand, follows a distribution closer to that of a Gaussian where there are no segments diverging from the general trend of the data. This can be graphically seen in the Gaussian fits to V in the right-hand subplots. The Gaussian fit to V is followed much tighter with LOFAR data as opposed to VLITE data and this provides a reason for why the anomaly detection strategy yielded poorer results for VLITE data compared to LOFAR data. As logistic regression does not require fits of thresholds to the $\eta - V$ 2D parameter space and works instead with greater number of parameters, this strategy works better at finding transient candidates in these VLITE data.

We additionally look into reasons for the existence of false positives in our results. These false positives represent stable sources that were incorrectly classified as transients and given the goal of prioritizing by precision, we would like to minimize the number of false positives (in red). The existence of false positives can be further analyzed by looking at these results in plots of η and V versus the maximum flux in Figure 8. This plot shows that the sources incorrectly identified as transients have a low maximum flux and is be subject to noise level sensitivity effects thereby leading to their misclassification.

7. *Conclusions and Future Work*

The use of machine learning strategies is an ever-increasingly useful tool to classify data. In this research, we aimed to find explosions in the universe by studying the variability or change in brightness of sources in VLITE images with injected simulated transients. Using anomaly detection and

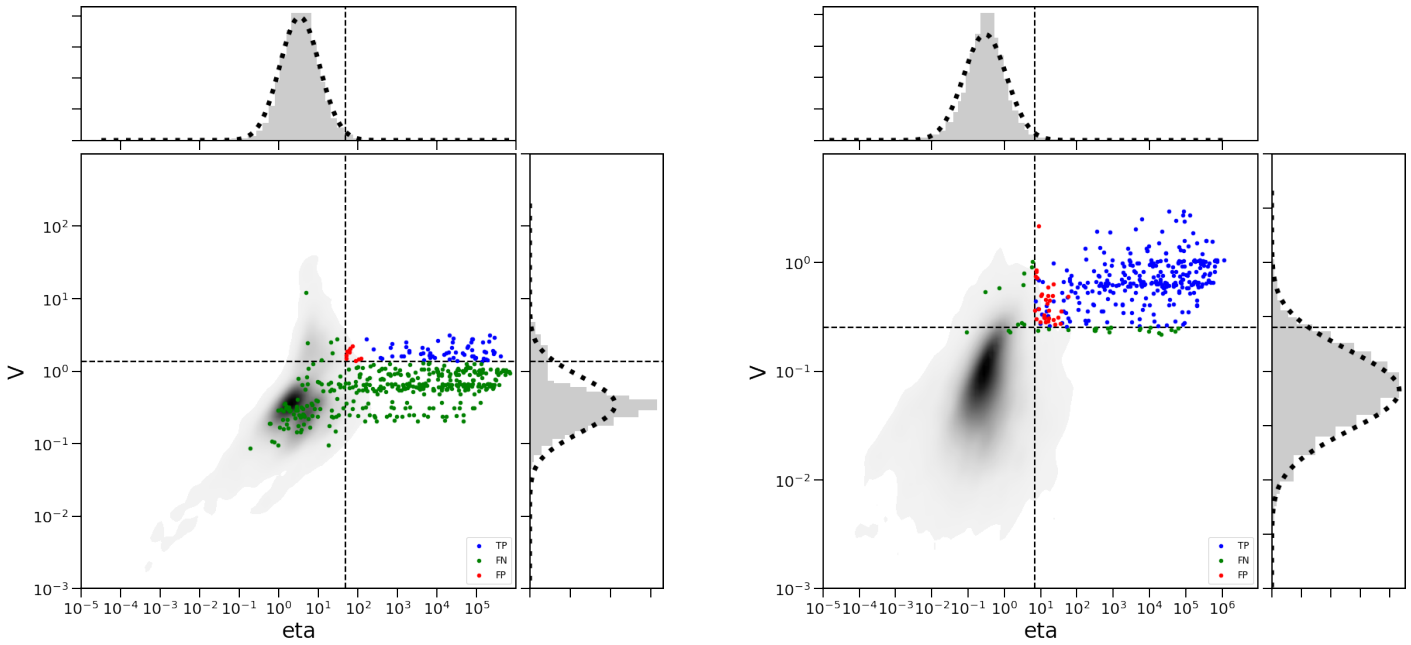

Figure 6: Results of the anomaly detection strategy for VLITE data with simulated injected transients (left) with a precision and recall of 0.84 and 0.14 respectively. This is compared to the anomaly detection results run against LOFAR data (right) with a precision and recall of 0.90 and 0.86 respectively.

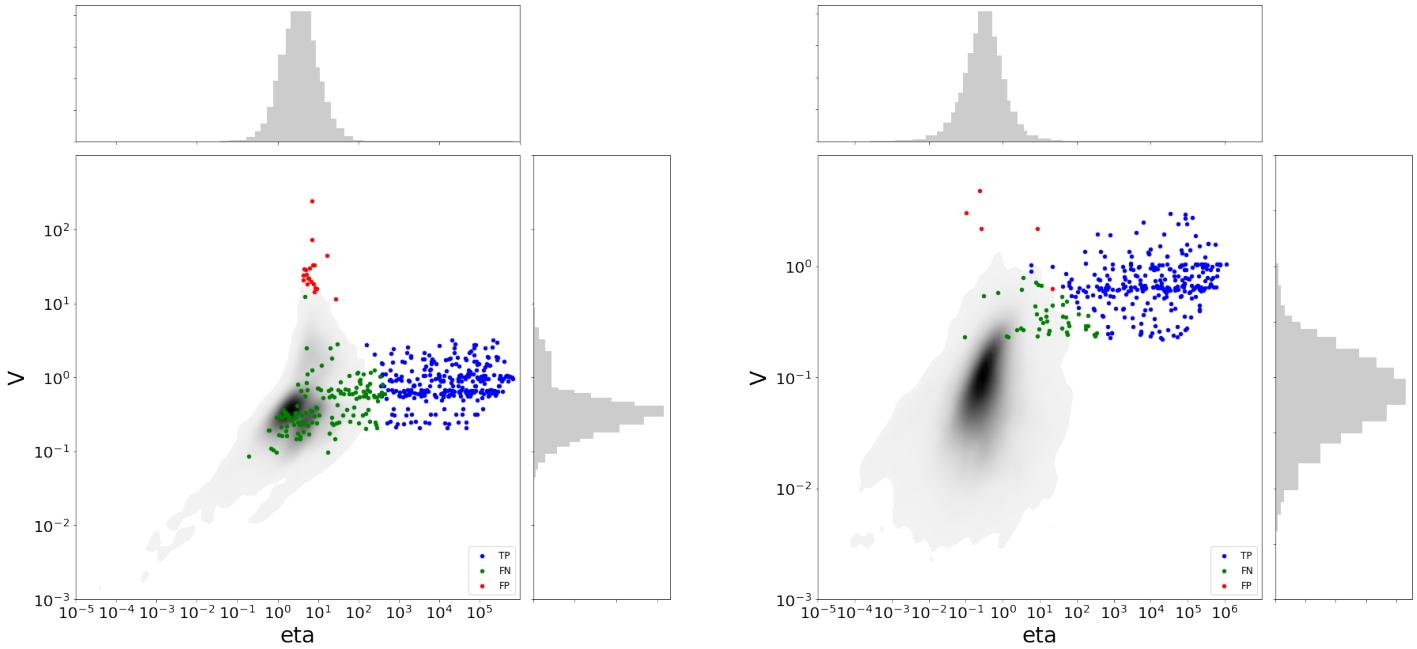

Figure 7: Results of the logistic regression strategy for VLITE simulated data (left) with a precision and recall of 0.93 and 0.68. Compared to the logistic regression results run of LOFAR data (right) with a precision and recall of 0.98 and 0.86.

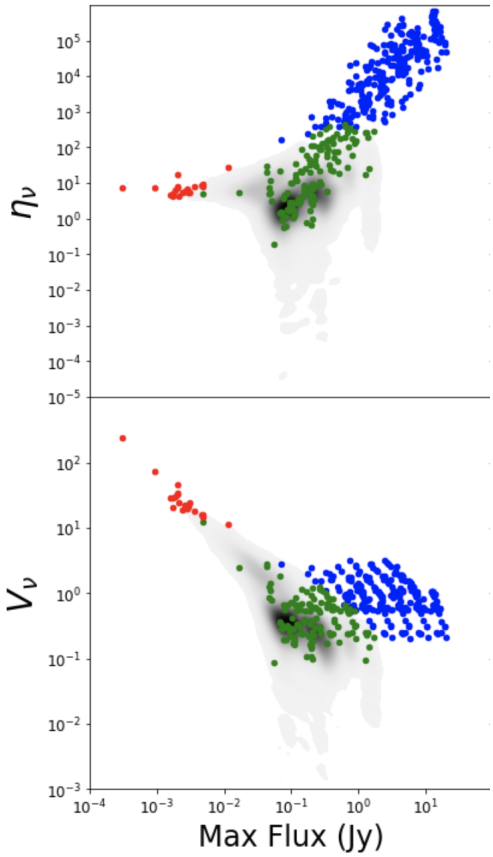

Figure 8: η and V versus Max Flux (Jy) displaying low maximum flux of false positive sources (red)

logistic regression machine learning algorithms, we found that the logistic regression strategy provided better precision and recall values. The non-Gaussianity of the distributions of η and particularly V in Figure 6 for VLITE data played a key reason for the underperformance of the anomaly detection strategy.

Additionally, we looked into levels of systematic uncertainties in flux measurements made by TraP. This is estimated using a `SciPy` Gaussian fit to the distribution of η when plotted against the peak flux and finding the value of smallest width. These Gaussian fits for varying systematic percentage levels showed the fit minimizing at $\approx 40\%$. We thus provide an estimate of $\approx 40\%$ systematic flux uncertainty for future VLITE transient searches.

Future work still to be conducted on this work include rerunning our machine learning strategies with η that accounts for systematic uncertainties. This will give us a better idea of the results of these codes when taking into account all uncertainties. Lastly, further work can be done in studying the source types to gain a fuller and more connected picture of what truly is behind these transients in VLITE data.

Acknowledgements

I would first like to thank my advisor, Alexander van der Horst, for his continued support, advice, and help throughout this project. I would also like to thank Antonia Rowlinson, Sarah Chastain, Alexandra Weikert, Emil Polisensky, and Caden Gobat for their assistance in this work. We would also like to thank the STEM Summer Research Program of the GW Columbian College of Arts and Sciences, the National Society of Physics Students, and the Luther Rice Fellowship for their support. This project would not be possible without the support of all mentioned above and my peers.

References

LOFAR. ASTRON, www.astron.nl/telescopes/lofar/.

Pietka, M., et al. "The Variability Timescales and Brightness Temperatures of Radio Flares from Stars to Supermassive Black Holes." ArXiv.org, 4 Nov. 2014, arxiv.org/abs/1411.1067.

Rowlinson, Antonia, et al. "Identifying Transient and Variable Sources in Radio Images." ArXiv.org, 18 Mar. 2019, arxiv.org/abs/1808.07781.

Swinbank, John D., et al. "The LOFAR Transients Pipeline." ArXiv.org, 5 Mar. 2015, arxiv.org/abs/1503.01526.

VLA Low Band Ionospheric and Transient Experiment (VLITE), vlite.nrao.edu/.

van Haarlem, M. P., et al. "LOFAR: The LOw-Frequency ARray." ArXiv.org, 19 May 2013, arxiv.org/abs/1305.3550.