Summer Internship Project Report

Exploring the Nature of Radio Sources using VLA Sky Survey

Submitted by

Satypriya Das
4^{th} year, MS in Astronomy and Astrophysics
(Dual Degree),IIST

Under the guidance of

Resmi Lekshmi
Associate Professor

Department of Earth and Space Sciences
INDIAN INSTITUTE OF SPACE SCIENCE AND TECHNOLOGY
THIRUVANANTHAPURAM, KERALA

Summer Internship 2023

Department of Earth and Space Sciences

INDIAN INSTITUTE OF SPACE SCIENCE AND TECHNOLOGY

Certificate

This is to certify that the internship report titled "Exploring the Nature of Radio Sources using VLA Sky Survey" submitted by Satyapriya Das, to the Indian Institute of Space Science and Technology, Thiruvananthapuram, is a bonafide record of the original work carried out by him under my supervision. The contents of this internship report, in full or in parts, have not been submitted to any other Institute or University.

Dr. Resmi Lekshmi
Department of Earth and Space Sciences
IIST

Place: Thiruvananthapuram
Date:

Abstract

The Very Large Array Sky Survey (VLASS) is a radio sky survey using the Karl G. Jansky Very Large Array (VLA). The survey covers the entire sky north of declination -40 degrees at frequencies of 2-4 GHz, with an angular resolution of 2.5 arcseconds. Gordon et al. paper on epoch 1 data is reproduced. Cleaning of both epoch 1 and 2 data is done based on that. Cross-catalog matching of VLASS is done with 2SXPS, Fermi-LAT, FIRST, GAIA, WISE, and SDSS to cover all frequency ranges. Using this data unsupervised (PCA,umap,t-sne, hdbscan)and supervised matching learning(random forest, SVM, xgboost, lightgbm) is done to classify BL Lac and Flat Spectrum Radio Quasar (FSRQ) from Fermi-LAT. Basic cross-epoch analysis is also done to find transients and in case of confusion in NED classification, images from VLA cutout are analysed and noted for future reference. In this project python, TOPCAT, and DS9 are used.

Contents

Chapter 1

Objective

Our interest is in point sources and we have primarily two aims.

- Classify the detected point sources based on their properties. Search the catalogue(s) for sources we expect to be there. For example:- positions of known supernovae are searched and in some cases, radio emission at the position is discovered.
- Identify variables/transients between epoch 1 and epoch 2.

So for that first, we have to understand how things are working hence we will also reconstruct the paper "A Quick Look at the 3 GHz Radio Sky. I. Source Statistics from the Very Large Array Sky Survey" from the tabular data (in CSV format) provided by the Canadian Initiative for Radio Astronomy Data Analysis (CIRADA). And based on this data further analysis will be done.

Chapter 2

Review and Recreation of "A Quick Look at the 3 GHz Radio Sky. I. Source Statistics from the Very Large Array Sky Survey"

2.1 Introduction

The Very Large Array Sky Survey (VLASS) is a synoptic radio sky survey using the Karl G. Jansky Very Large Array (VLA). The survey covers the entire sky north of declination -40 degrees at frequencies of 2-4 GHz, with an angular resolution of 2.5 arcseconds. The VLASS is the largest and most sensitive radio survey ever conducted, and it is expected to catalog over 10 million radio sources.

The VLASS is a powerful tool for studying the evolution of the universe. It can be used to study the formation and evolution of galaxies, the growth of supermassive black holes, and the distribution of cosmic gas. The VLASS is also expected to discover new and unexpected phenomena, such as transient radio sources and new types of galaxies.

Based on Quick Look images from this first epoch, CIRADA have created a catalog of 1.9×10^6 reliably detected radio components. Due to the limitations of the Quick Look images, component flux densities are underestimated by 15% at Speak > 3 mJy $beam^{-1}$ and are often unreliable for fainter components. We use this catalog to perform statistical analyses of the ν around 3 GHz radio sky. Comparisons with the Faint Images of the Radio Sky at Twenty Centimeters (FIRST) survey show the typical 1.4–3 GHz spectral index, α, to be around 0.71. The radio color–color distribution of point and extended components is explored by matching with FIRST and the LOFAR Two-meter Sky Survey. We present the VLASS source counts, dN/dS, which are found to be consistent with previous observations at 1.4 and 3 GHz. Resolution improvements over FIRST result in excess power in the VLASS two-point correlation function at angular scales 7 arcsecs, and in 18% of active galactic nuclei associated with a single FIRST component is split into multi-component sources by VLASS.

2.2 Key points regarding VLASS data from the research paper

Here the analysis of the epoch 1 images is done. For each of the 35,285 subtitles in VLASS, the source extraction code Python Blob Detection and Source Finding (PyBDSF; Mohan and Rafferty 2015) was run in "srl" mode— meaning that a list of components is provided rather than a list of individual Gaussian fits. The method adopted by PyBDSF is to detect flux islands at 3σ above the image mean (where σ is the local rms) and then fit components composed of one or more Gaussians within those islands with a peak brightness at 5σ above the image mean.

These sources are then stored in the CSV file (uncleaned) which is available for use. There are various flags that they have used to clean the data. They are as follows:

- **Peak_to_ring:** The ratio of peak brightness to the maximum flux density in a ring centered on the component position and with inner and outer radii of 5″ and 10″, respectively. This is specifically designed to identify potential sidelobe structures that have erroneously been fitted as components.
- **Duplicate_flag:** 0 = unique, 1 = duplicate but preferred,2 = duplicate and not preferred.
- **Quality_flag:** $S_{peak} > S_{total}$ by setting Quality_flag = Quality_flag+4. Flag components with a peak brightness lower than 5 times the local rms and set Quality_flag = Quality_flag +2. This metric is only used to flag components that are more than 20″ from another component and with Peak_to_ring < 2 hence Quality_flag = Quality_flag + 1. Quality_flag == 0, indicating that the component has not been flagged for any of the above quality issues, in addition to just applying simple brightness cuts to the data. In addition to that maximum of Quality_flag == 4 found in real sources by random sampling hence they are also included in further analysis.
- **S_Code:** S = single-Gaussian component that is the only component in that particular flux island. C = single-Gaussian component in a flux island with other components. M = multi-Gaussian component. E = empty flux island, i.e. a flux island was detected by PyBDSF, but no components were fitted.

  Hence S_Code != E is used as a constraint since it is empty and only considered for the sake of completeness.

From all these, the constraint formed is:
Duplicate_flag < 2 , Quality_flag = 0 or 4 , S_Code ! = 'E'

Based on comparisons with > 50 VLA calibrator sources it is estimated that the peak brightness in the VLASS Quick Look images is underestimated by around 15%, and the total flux density of components is underestimated by around 10%.

The data quality issues such as:

- At the southern declinations a "checkerboard" pattern of rms is indicative of variation between VLASS tiles and sub-epochs.
- At northern latitudes, in particular, high noise levels are visible at the east–west boundaries of VLASS tiles. The increase in noise is likely to the fact that some tile edges were flagged to avoid unreliable fluxes associated with online software bugs related to the ghosts.
- There is a region of high noise that stands out at declination around +85° and 12 hr < R.A. < 24 hr.
- The Galactic plane is clearly visible as a region of enhanced noise, albeit only at very low Galactic latitudes ($\|b\| < 0.5°$) and for a relatively narrow range of Galactic longitudes ($350° < l < 50°$).
- There are several strips at $0° < declination < +1°$ that have lower noise (around 100 Jy $beam^1$). This is the result of accidentally observing these regions twice during survey operations and using both observations in the production of the Quick Look images. This deeper VLASS region has a total area of around 160 deg^2. .

2.3 My results

The data of epoch 1 and 2 which is available in the csv format available in CIRADA website is downloaded. Cleaning is done by applying the constraints in Python based on the steps mentioned in the current paper. The file is opened as a pandas dataframe (named VLA). Then the constraint was applied as:

Constraint = VLA.loc[(VLA['S_Code'] != 'E') & (VLA['Duplicate_flag'] < 2) & (VLA['Quality_flag'].isin([0, 4]))]

Applying this to both epochs 1 and 2 gave 1,739,566 and 1,873,524 sources respectively. This number is a bit less than the given number in the paper but does not affect much since it is on the lower side rather than the higher side rechecking is done a few times but these numbers are self-consistent.

After that astrometry is done by cross-catalog analysis of epoch 1 with GAIA DR3 (in the Gordon paper it is done with GAIA DR2). Using that Fig. 2.1 and 2.2 are plotted. As we can see in Fig. 2.1 that the ΔRA average is around 0 arcsec but ΔDEC is offset and peaks around -0.25 arcsec. This thing is also visually more clear in Fig 2.2.

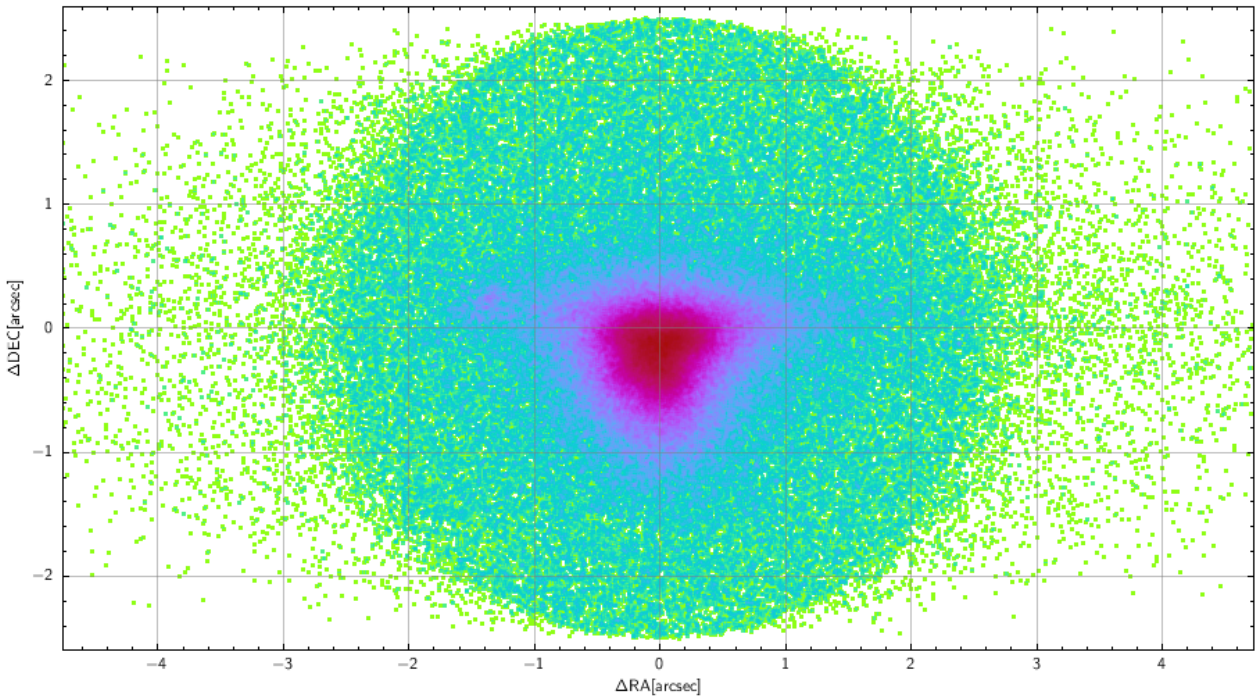


Figure 2.1: ΔRA ($RA_{GAIA} - RA_{VLASS}$) vs ΔDEC ($DEC_{GAIA} - DEC_{VLASS}$)

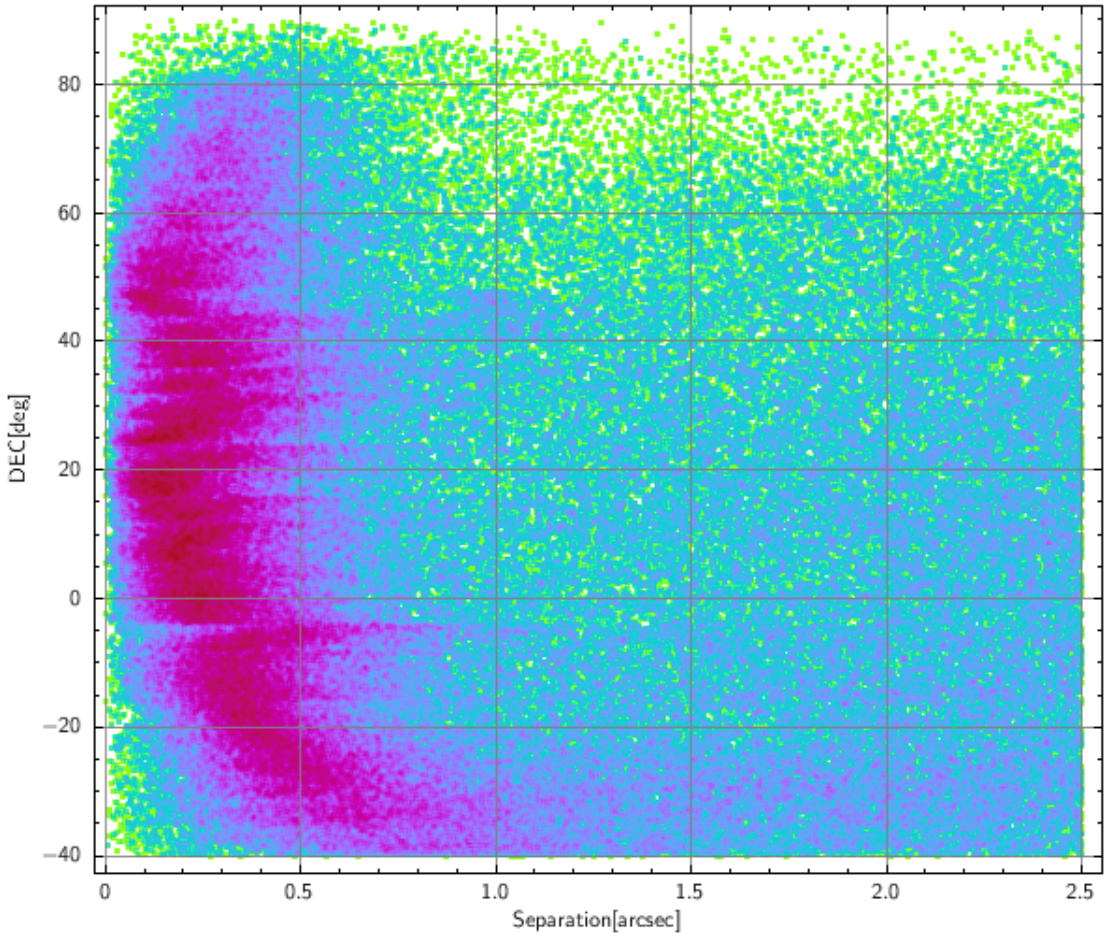


Figure 2.2: Separation(VLASS and GAIA) vs DEC (VLASS)

Fig 2.3 which I have reproduced is used to show the resolved and unresolved sources. The peak_flux<3 represents resolved and above 3 represents unresolved hence to have a good analysis it will be optimal to take unresolved sources for analysis. The color depicts the density of points in the figure and the vertical line at peak_flux ==3 separates the graph into two regions as explained above.

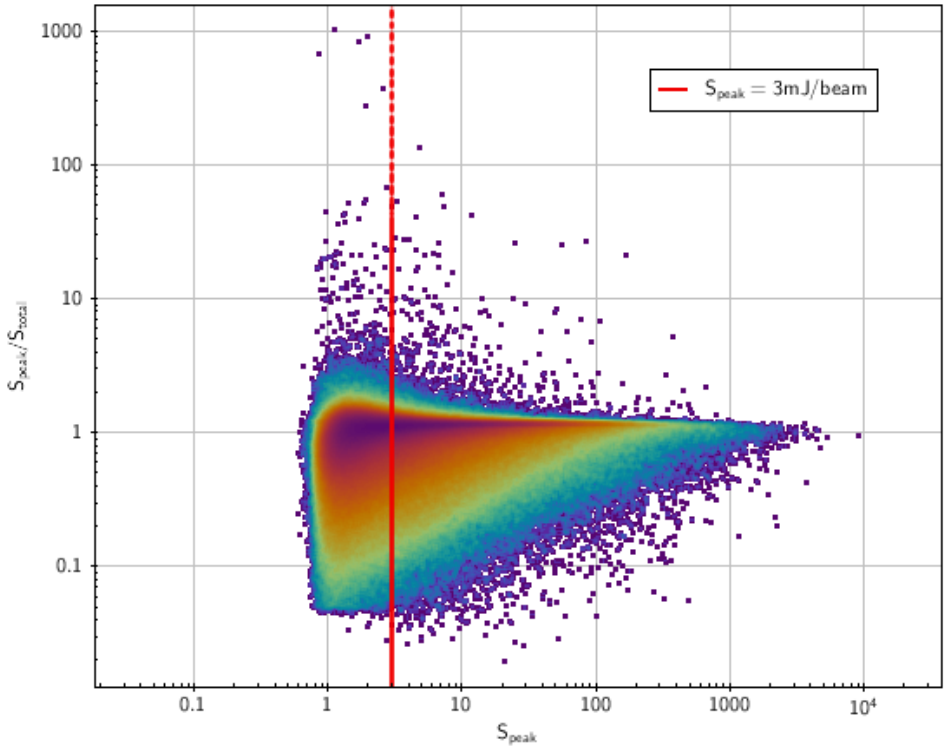


Figure 2.3: The ratio of peak brightness to total flux density as a function of peak brightness after constraint

Fig 2.2 shows the relation between radio source size and total flux density. VLASS is run in the VLA's B and BnA configurations. One of the impacts of this observing mode is that sensitivity to large angular scale structures is reduced by a lack of short baselines used in the array configuration. VLASS results in larger (ψ around 30″) objects being resolved into multiple components or lost altogether. Consequently, components detected in a high angular resolution such as VLASS are only suitable for testing this relation at small angular scales, and testing at large angular scales necessitates a sample of verified multi-component sources.

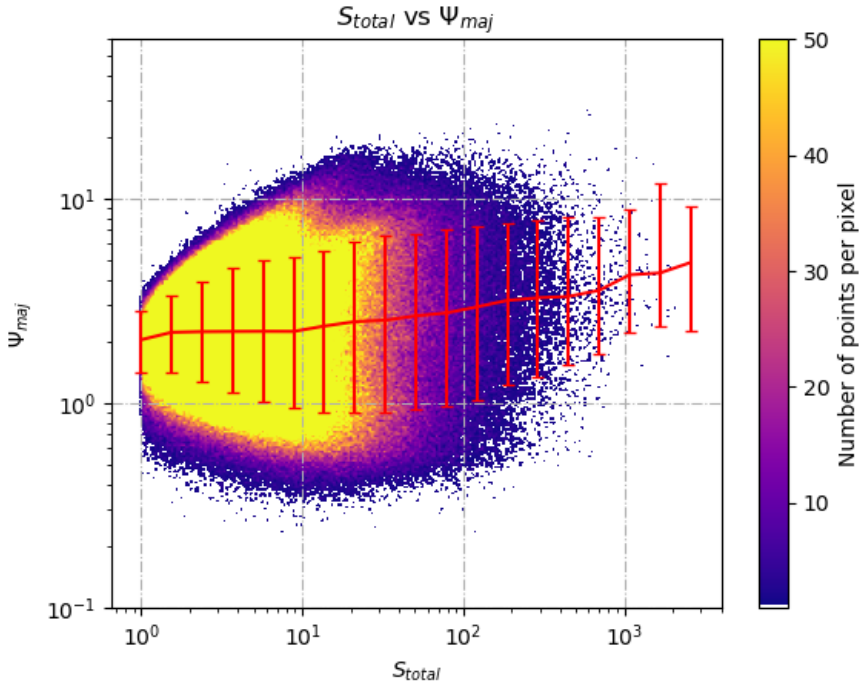


Figure 2.4: The distribution of the deconvolved major axis for VLASS components with a nonzero cataloged size vs. component total flux density. The red line shows the median deconvolved angular size of the component for a particular flux bin. Error bars are defined by the 16^{th} and 84^{th} percentiles of the size distribution in each flux bin

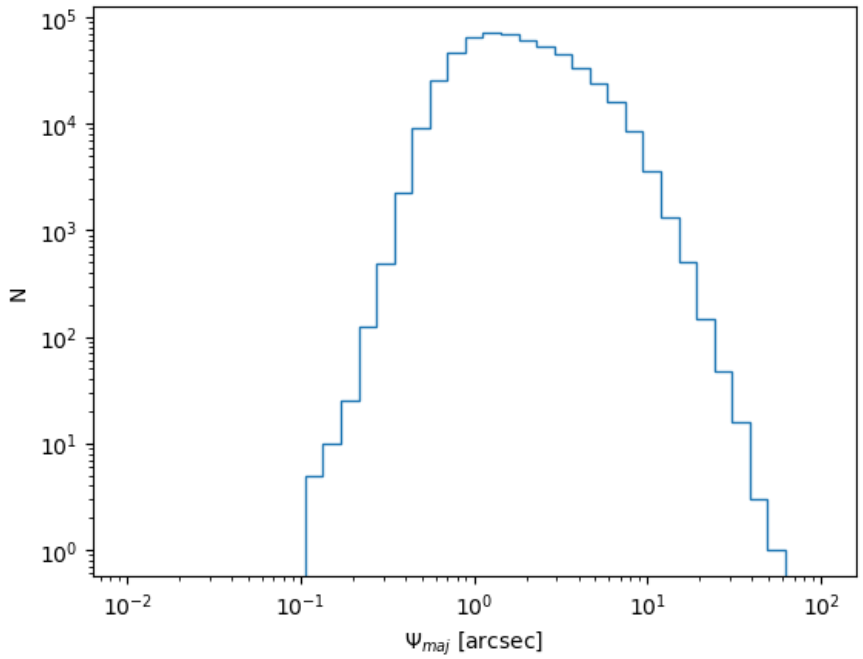


Figure 2.5: Histogram of the deconvolved major axis size, ψ_{maj}

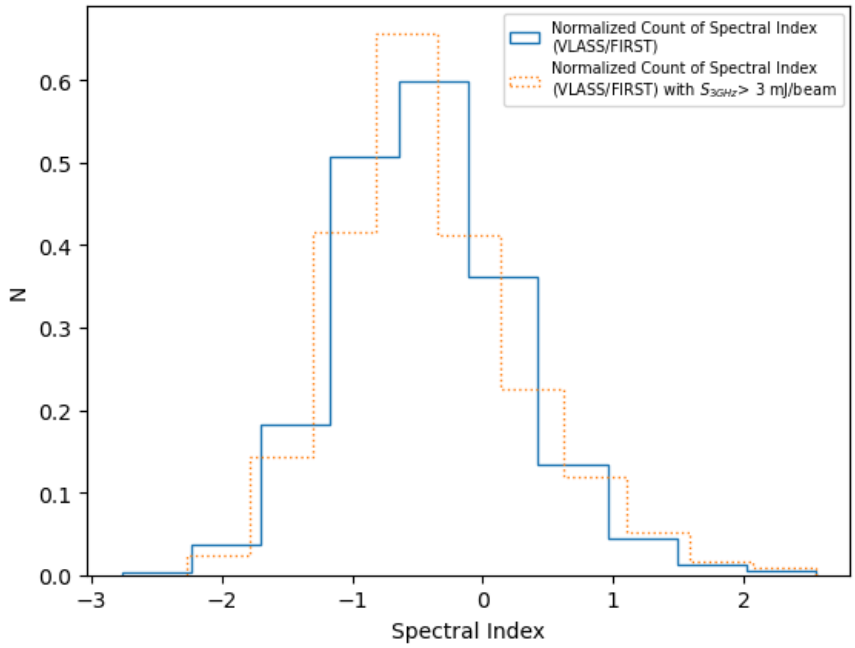


Figure 2.6: Normalized distributions of the spectral index, α, for VLASS components associated with FIRST components.

Cross-catalog analysis is done for the entire catalog and Landy & Szalay (1993) estimator is used to find a two-point correlation as discussed in the paper review section. Doing all this is possible but that will require a system with high computational power, which is beyond the capacity of the resources I have (my laptop and the institute's GPU cluster). Hence the results are taken for granted and used as it is if required. With the cleaned data further analysis can be done which is mentioned/used in the subsequent chapters.

Chapter 3

Blazar Classification using Multifrequency Data

3.1 Introduction

3.1.1 Blazars

Blazars are among the most enigmatic and energetic objects in the universe, they are special subclass of active galactic nuclei (AGN) characterized by their highly variable emission and relativistic jets pointed nearly in the direction of Earth. These jets, composed of ionized matter traveling at nearly the speed of light, produce intense radiation across the electromagnetic spectrum, ranging from radio waves to gamma rays. Thus, it is part of our studies.

Key Characteristics of Blazars:

- **Relativistic Jets:-** Blazars are distinguished by their relativistic jets, which emit radiation that appears much brighter due to relativistic beaming, an effect caused by the high speeds of the jet particles. Thus they have high radio and optical polarization.
- **Variability:-** Blazars are highly variable sources, exhibiting rapid and dramatic fluctuations in brightness on timescales as short as hours to days.
- **Emission Across the Spectrum:-** Blazars emit radiation across the entire electromagnetic spectrum, thus it has a broad continuum extending from the radio through the gamma rays.
- **Power Source:-** The energy source for blazars is believed to be a supermassive black hole at the center of their host galaxy. As material falls into the black hole, it forms an accretion disk, which heats up and releases enormous amounts of energy. This is why its signal has core-dominated radio morphology, flat radio spectra, and apparent superluminal motion.

Blazars are usually divided into flat-spectrum radio quasars (FSRQs) and BL Lacertae objects (BL Lacs) according to the equivalent width (EW) of their optical emission lines. But here we will try to classify based on the multi-frequency data acquired from various catalogs.

3.1.2 Machine Learning

Machine learning is an emergent field in the modern era that has immense application and can be used to find the underlying relation between various data obtained for a given problem. Here, its application in astronomy via the blazar classification is realized.

The learning system of a machine learning algorithm into three main parts.

- **A Decision Process:-** In general, machine learning algorithms are used to make a prediction or classification. Based on some input data, which can be labeled or unlabeled, your algorithm will produce an estimate of a pattern in the data.

- **An Error Function:-** An error function evaluates the prediction of the model. If there are known examples, an error function can make a comparison to assess the accuracy of the model.
- **A Model Optimization Process:-** If the model can fit better to the data points in the training set, then weights are adjusted to reduce the discrepancy between the known example and the model estimate. The algorithm will repeat this "evaluate and optimize" process, updating weights autonomously until a threshold of accuracy has been met.

It is further classified into the following:-

- **Supervised machine learning:-** It uses labeled datasets to train algorithms to classify data or predict outcomes accurately. As input data is fed into the model, the model adjusts its weights until it has been fitted appropriately. This occurs as part of the cross-validation process to ensure the model avoids overfitting or underfitting. Some examples are random forest, support vector machine (SVM), XgBoost, LightGBM, neural network, etc.
- **Unsupervised machine learning:-** It uses machine learning algorithms to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention. This method's ability to discover similarities and differences in information makes it ideal for exploratory data analysis etc. It's also used to reduce the number of features in a model through the process of dimensionality reduction. Principal component analysis (PCA), umap, and t-sne are some examples of it.

3.2 Data collection and combining for ML

Catalogs used:-

- Fermi LAT 14-year point source catalog
- VLASS epoch 1 catalog
- 2SXPS catalog
- GAIA AGN Catalog (GLEAN)
- SDSS Quasar Catalog
- WISE AGN Candidate Catalog

Matching Algorithm used:- Sky with ellipses,cross-match in TOPCAT

Here in this analysis, the blazar classified data from Fermi-LAT is used. As we know the resolution in the gamma region is low. It is limited by non-gamma-ray backgrounds at lower energies, and, at higher energy, by the number of photons that can be detected. Hence direct match of Fermi-LAT with catalogs can't be done. Hence the error ellipse is considered while doing cross-match with it. Those Fermi-LAT sources which has error ellipse data, have significance>4 and belong to BL Lac (BLL), Flat Spectrum Radio Quasar(FSRQ), Blazar Candidate of Unknown Type (BCU) are selected for further analysis. Its cross-matched with 2SXPS is done by considering the best match to the Fermi ellipse. It is then used as a pivot point to do cross-matching with other catalogs. Other catalogs are also shortlisted by doing one-to-many ellipse cross-match (to get all those inside the Fermi ellipse). All these catalogs are then cross-matched all at once using the quadrature cross-match feature of TOPCAT (matching 5 tables at a time).

After this, those columns with photometric data are selected. Since the distance from the source is not available (GAIA parallax takes positive and negative values for the overall analysis of objects, since it was doubtful it is not used). Hence to get distance-independent parameters, colors are calculated using the magnitudes in each band. The analysis of these columns is done by plotting the normalized histogram for both classes i.e. FSRQ and BLL so that can be used in machine learning.

Table 3.1: Astrometric Data used For Cross-matching

Catalog	Column Name	Count	Description
VLASS	RA	972	Right Ascension
	Dec	972	Declination
	Dec_Maj	972	Deconvolved Major
	Dec_Min	972	Deconvolved Minor
2SXPS	RA	972	Right Ascension
	Dec	972	Declination
	Err90	972	Error radius
GAIA	RA	797	Right Ascension
	Dec	797	Declination
	RA_error	797	Right Ascension error(in arcsec)
	Dec_error	797	Declination error(in arcsec)
Fermi LAT	ra_combined	972	Combined data - Main RA if firm association otherwise associated RA
	dec_combined	972	Combined data - Main Dec if firm association otherwise associated Dec
	Semi_Maj_new	972	Combined data - Main Semi_Maj_95 if firm association otherwise associated error
	Semi_Min_new	972	Combined data - Main Semi_Min_95 if firm association otherwise associated error
WISE AGN Catalog	WISEA	417	WISE Name
	RA_WISE	417	Right Ascension
	Dec_WISE	417	Declination
SDSS Quasar Catalog	SDSS_NAME	140	SDSS Name
	RA_SDSS	140	Right Ascension
	Dec_SDSS	140	Declination

Table 3.2: Combined Multifrquency data for ML

Catalog	Column Name	Count	Description
VLASS	Peak_flux	972	Peak flux of the source
2SXPS	Hardness ratio-1	972	The HR1 hardness ratio = (M-S)/(M+S) where M and S are the medium (1-2 keV) and soft (0.3-1 keV) band count rates.
	Hardness ratio-2(HR2)	972	The HR2 hardness ratio = (H-M)/(H+M) where H and M are the hard (2-10 keV) and medium (1-2 keV) band count rates.
GAIA	phot_g_mean_mag	797	G-band mean magnitude
	bp_rp	797	BP-RP Color
	qso_variability	797	Quasar variability metric in the G band
	fractional_variability_g	797	Fractional variability in the G band
Fermi LAT	flux_density	972	Combing flux density from different spectral type
	spectral_index	972	Combing spectral index from different spectral type
	variability_index	972	Variability index

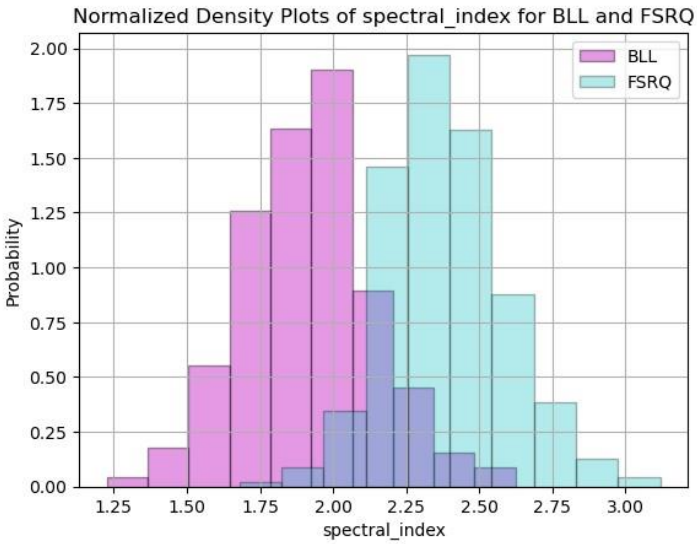
Continued on next page

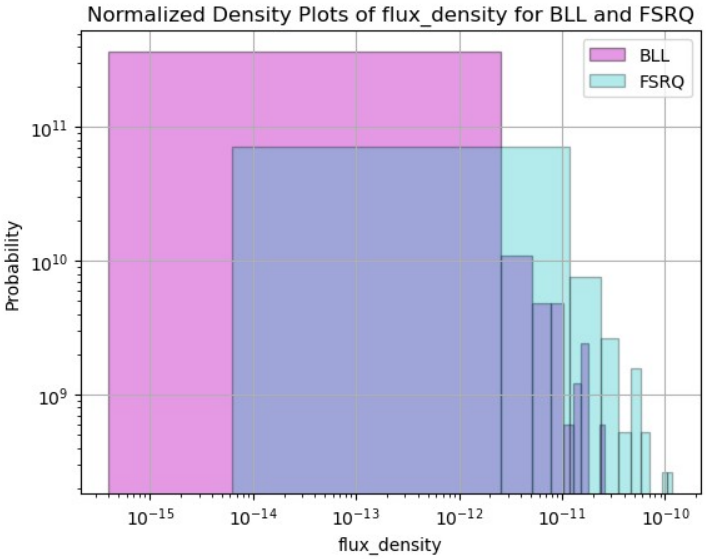
Table 3.2 – Continued from previous page

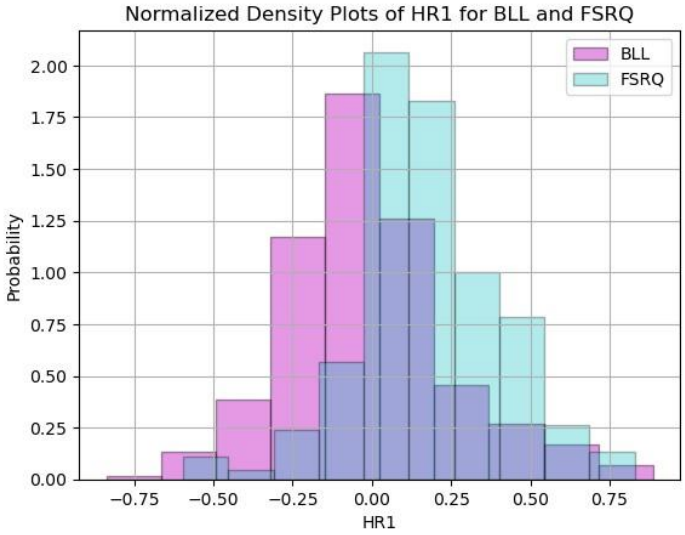
Catalog	Column Name	Count	Description
WISE AGN Catalog	W2mag_W3mag	417	(W2 magnitude - W3 magnitude) color
	W1mag_W4mag	417	(W1 magnitude - W4 magnitude) color
SDSS Quasar Catalog [REMOVED after ANALYSIS]	FUV_NUV	140	(Far_UV - Near_UV) color
	JMAG_KMAG	140	(JMAG - KMAG) color
	JMAG_HMAG	140	(JMAG - HMAG) color
	HMAG_KMAG	140	(HMAG - KMAG) color
For classification	source_type	972	Type of the source from Fermi-LAT

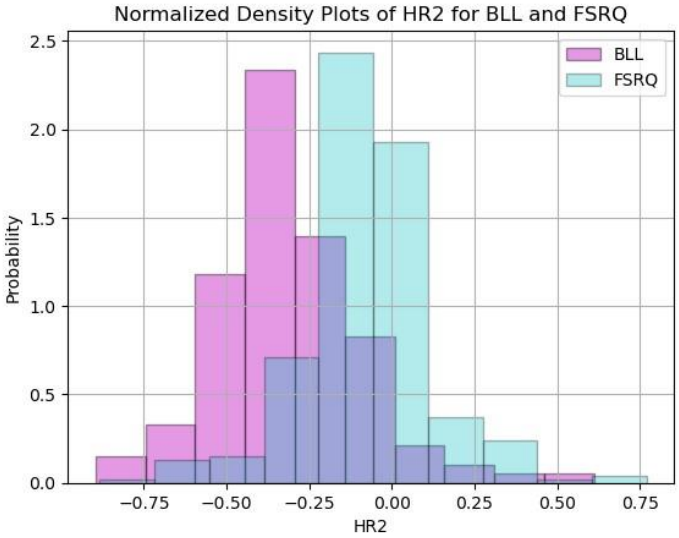
Now in order to further understand the importance of these columns in the classification, their normalized histogram are plotted. They are mentioned in the subsequent part.

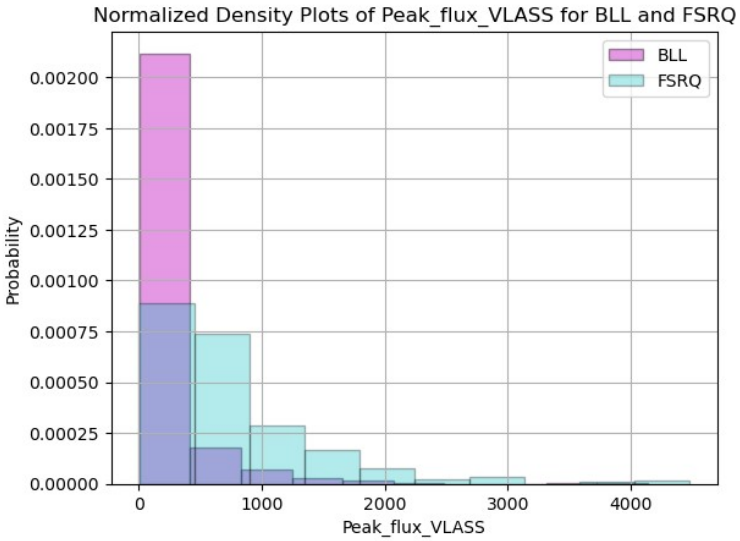
An iterative approach for analysis is used. From the basic list of useful columns. The SDSS data has low row count and the separation between classes was not there. Also out of 143, only 3 will be there in BCU making the analysis while predicting useless. Removing them reduced accuracy by 0.2% only but led to an increase in correct classification in the confusion matrix for each class by 1%. Hence it was a fair trade-off. For the WISE colors, these two had the maximum separation among other combinations. Hence only these two are used. Given columns are the derived columns from the various catalogs in order to find meaningful and useful data for the classification.

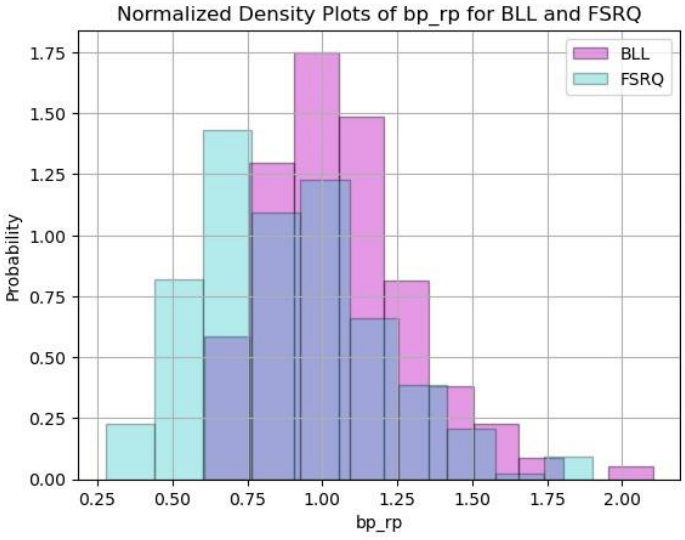

Normalized Density Plots of spectral_index for BLL and FSRQ
BLL
FSRQ
Probability
spectral_index

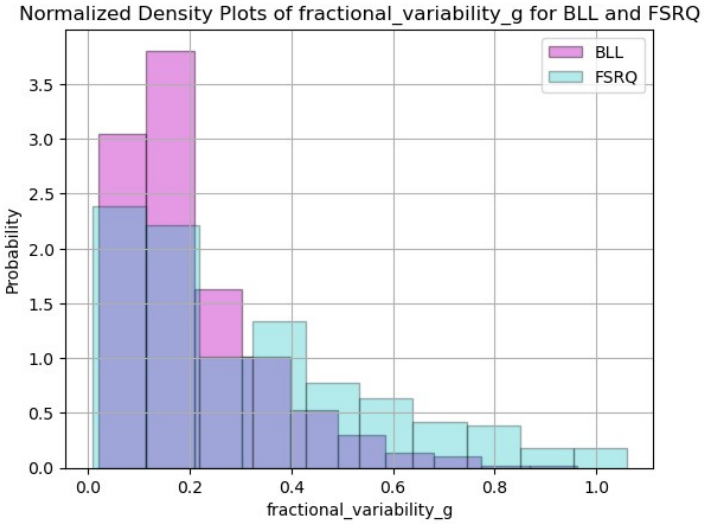

Normalized Density Plots of flux_density for BLL and FSRQ
BLL
FSRQ
Probability
flux_density

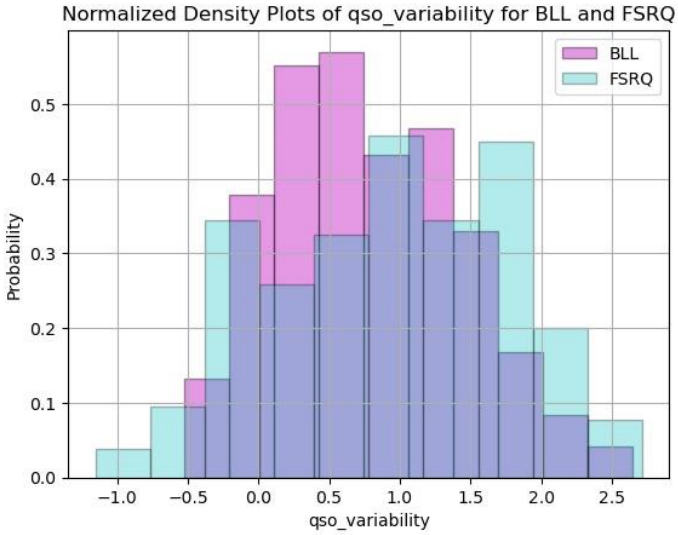

Normalized Density Plots of HR1 for BLL and FSRQ
BLL
FSRQ
Probability
HR1

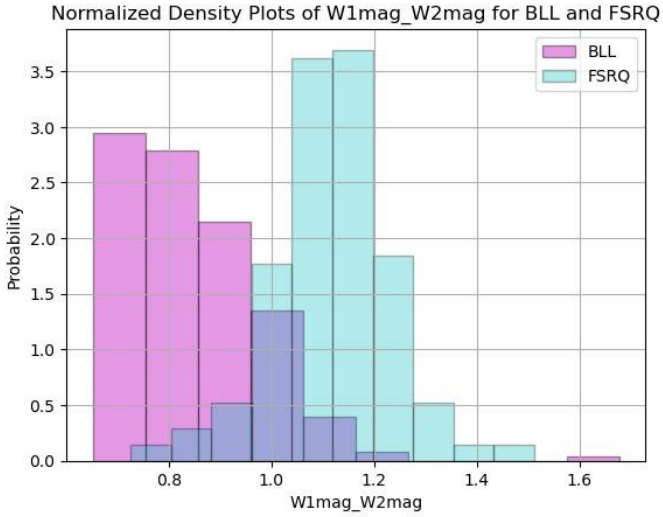

Normalized Density Plots of HR2 for BLL and FSRQ
BLL
FSRQ
Probability
HR2

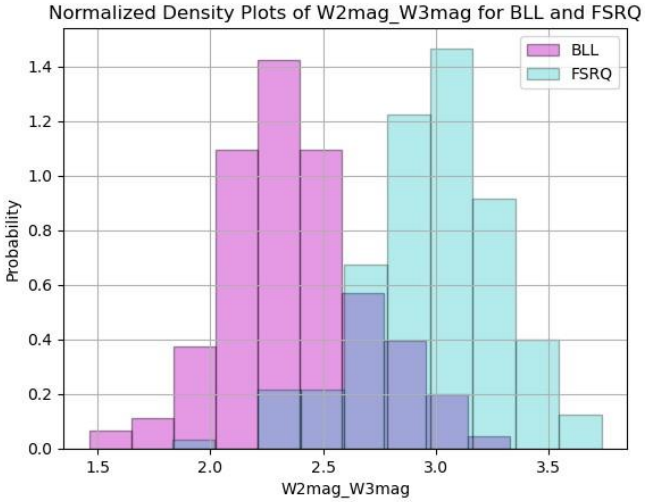

Normalized Density Plots of Peak_flux_VLASS for BLL and FSRQ
BLL
FSRQ
Probability
Peak_flux_VLASS

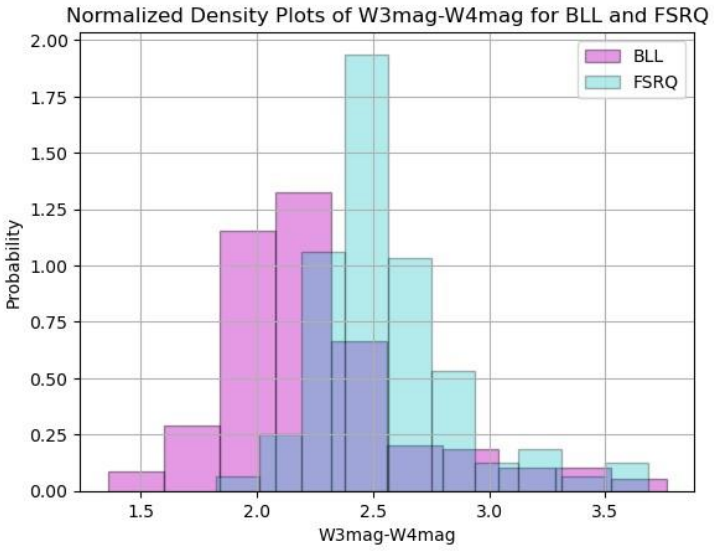

Normalized Density Plots of bp_rp for BLL and FSRQ
BLL
FSRQ
Probability
bp_rp

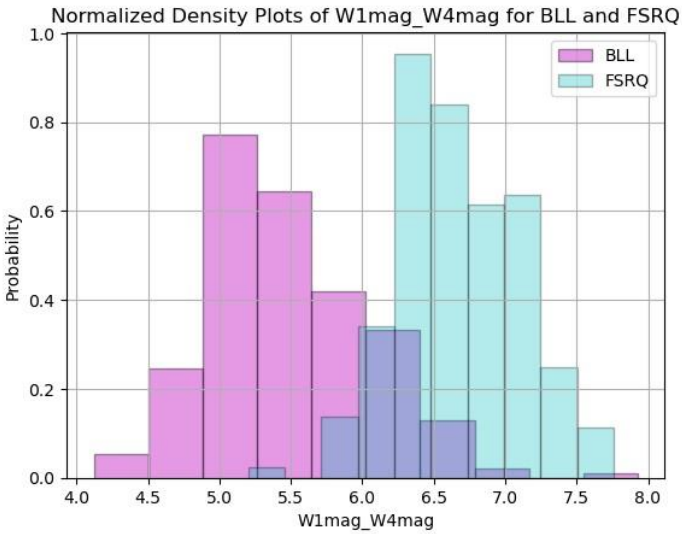

Normalized Density Plots of fractional_variability_g for BLL and FSRQ
BLL
FSRQ
Probability
fractional_variability_g

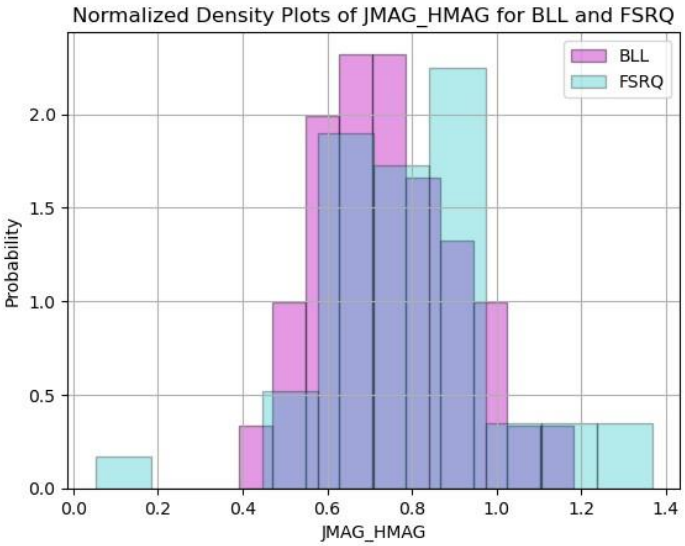

Normalized Density Plots of qso_variability for BLL and FSRQ
BLL
FSRQ
Probability
qso_variability

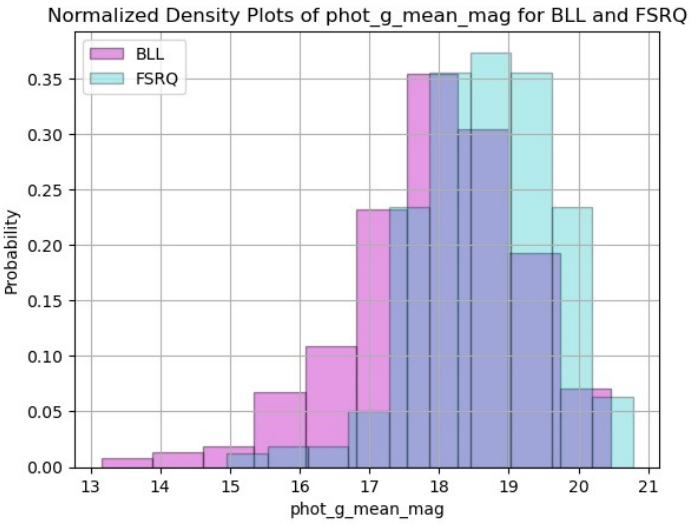

Normalized Density Plots of W1mag_W2mag for BLL and FSRQ
BLL
FSRQ
Probability
W1mag_W2mag


Normalized Density Plots of W2mag_W3mag for BLL and FSRQ
BLL
FSRQ
Probability
W2mag_W3mag

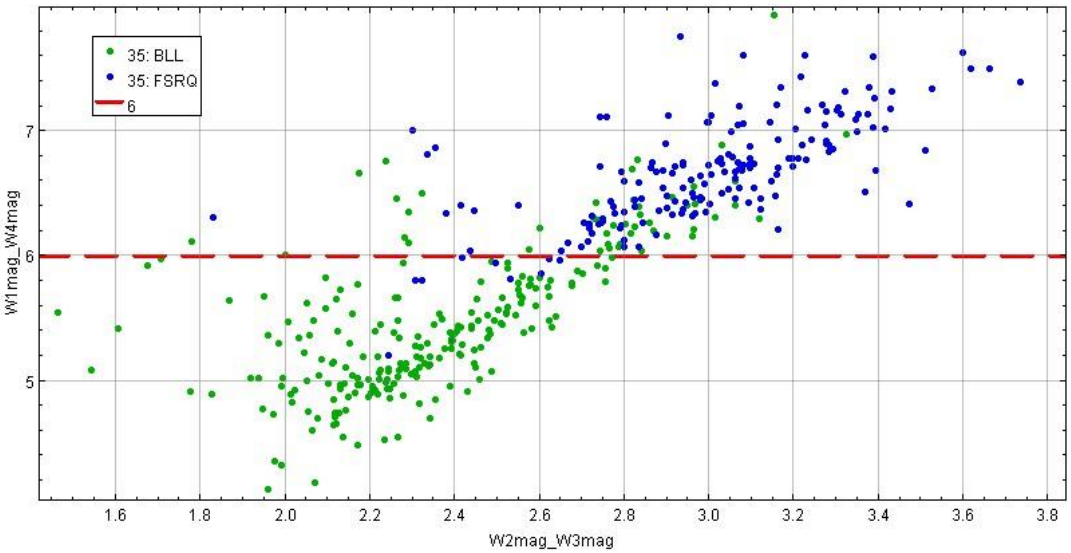

Normalized Density Plots of W3mag-W4mag for BLL and FSRQ
BLL
FSRQ
Probability
W3mag-W4mag


Normalized Density Plots of W1mag_W4mag for BLL and FSRQ
BLL
FSRQ
Probability
W1mag_W4mag





Color-color plot using WISE Data



3.3 Unsupervised Machine Learning

Here in unsupervised machine learning we use the data of given columns and try to find the clustering of the data for categorization. The data is imputed with the median of the columns. It is then normalized i.e. for each column the mean is zero and the standard deviation is 1. This removes the correlation between columns and makes them independent. Due to this, the model will give importance to all the columns equally with no partiality. This is a crucial pre-processing step that should be done before any kind of analysis be it supervised or unsupervised machine learning.

Various unsupervised machine learning techniques are used here. They are as follows:-

- **Principle Component Analysis (PCA):-** It is a dimensionality reduction technique that projects the most important data from the higher dimension to the lower dimension. It finds an orthogonal axis that covers most variance (principal component). For visual purposes, we will take only 2 principal components for analysis. It works by the first standardization that we did just now in pre-processing. Then it finds the covariance matrix to find the relation of each column with one another. This matrix is then decomposed into eigenvalues and eigenvectors. Eigenvectors are then selected based on the descending order of their eigenvalue. This is a very simple yet very powerful method.

- **UMAP:-** It is an algorithm for dimension reduction based on manifold learning techniques and ideas from topological data analysis. It provides a very general framework for approaching manifold learning and dimension reduction, but can also provide specific concrete realizations. It is implemented using the UMAP module of Python.

- **t-SNE (t-distributed Stochastic Neighbor Embedding):** It is a non-linear dimensionality reduction technique where we try to first find the SNE then approximate it using the student t-distribution. It is implemented by importing TSNE from sklearn.manifold .

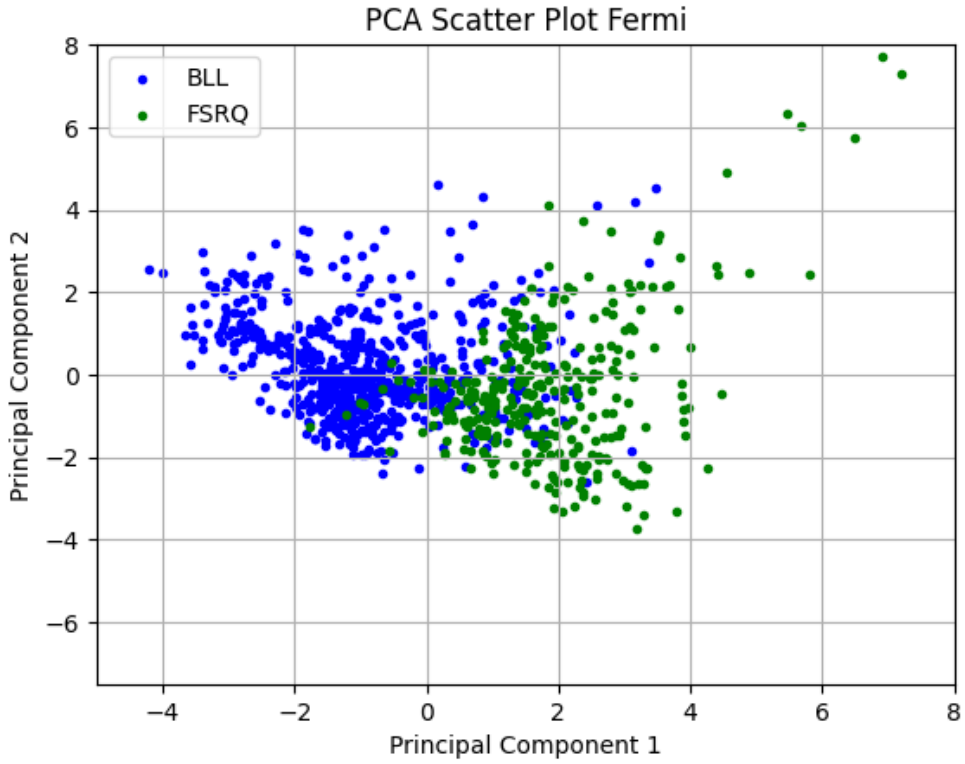


Figure 3.1: PCA of Multi-frequency analysis with labels from Fermi-LAT catalog

HDBSCAN is giving garbage results for all kinds of hyperparameters. Hence it can't be classified with that.

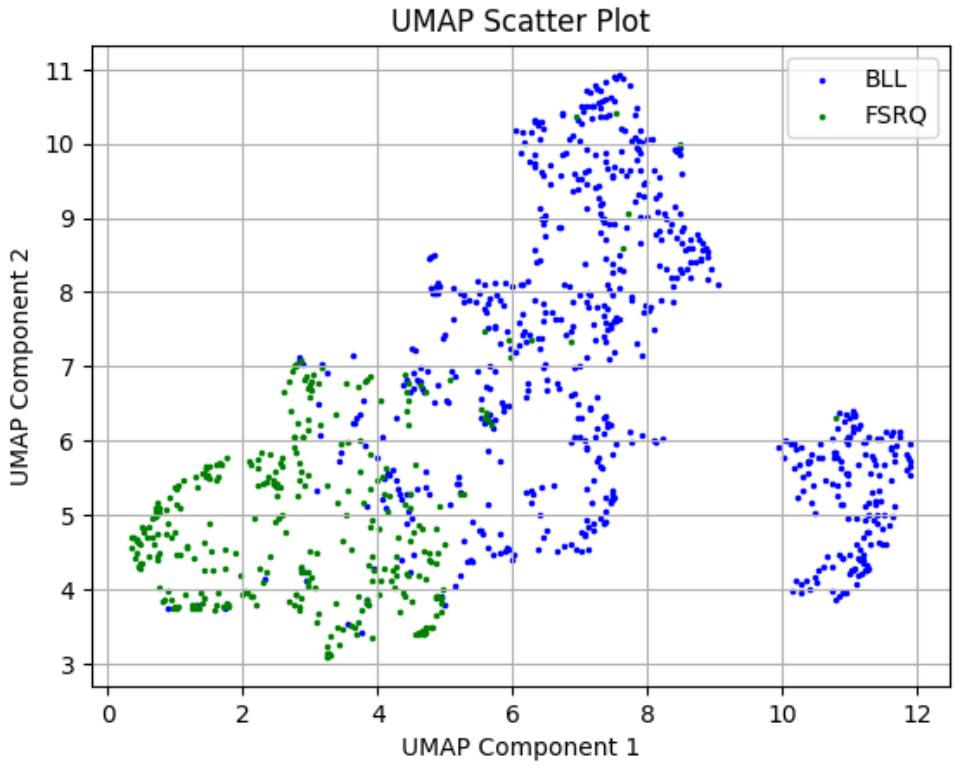


Figure 3.2: UMAP scatter plot of multi-frequency analysis with labels from Fermi-LAT catalog

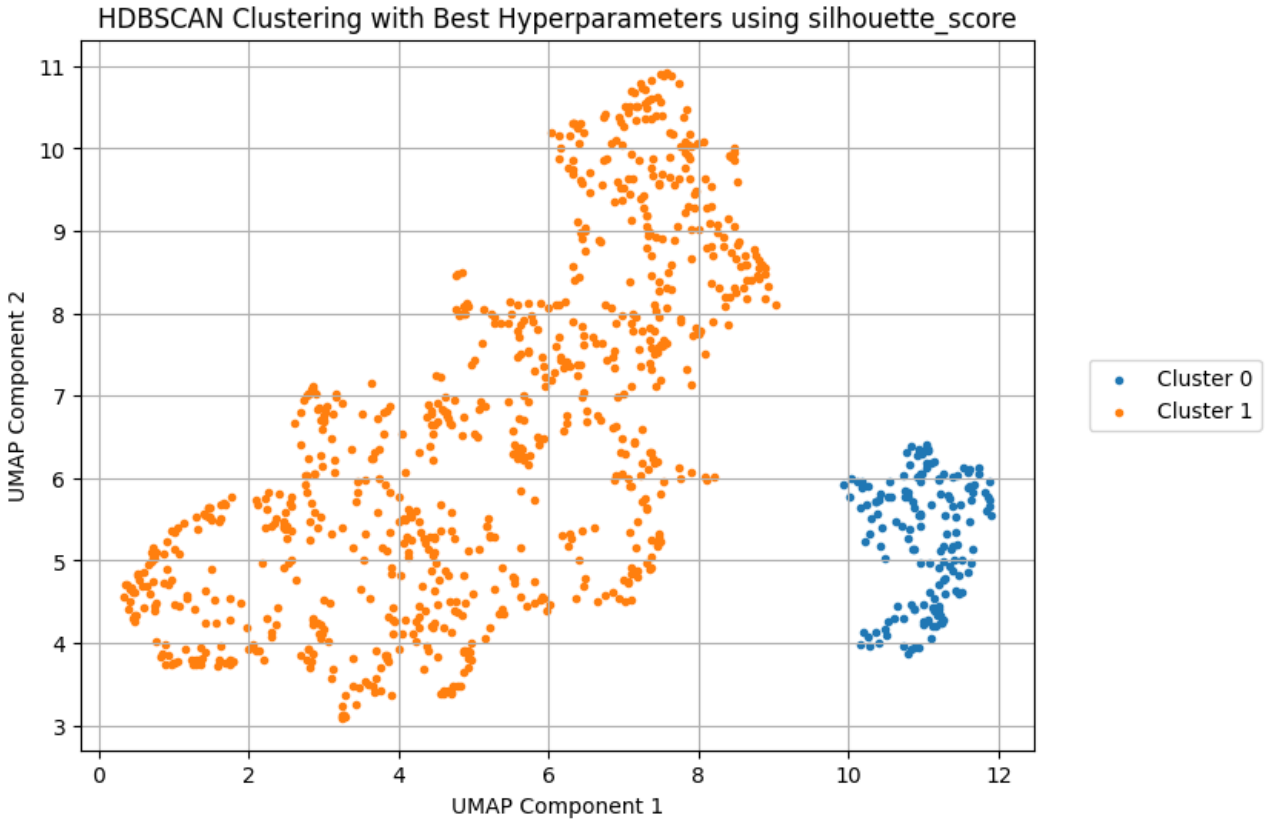


Figure 3.3: Clustering using hdbscan of a umap scatter plot of multi-frequency analysis using silhouette_score only. Best Hyperparameters are $min\_cluster\_size = 5$, $min\_samples = 3$, $cluster\_selection\_epsilon = 0$, Best Silhouette Score = 0.45

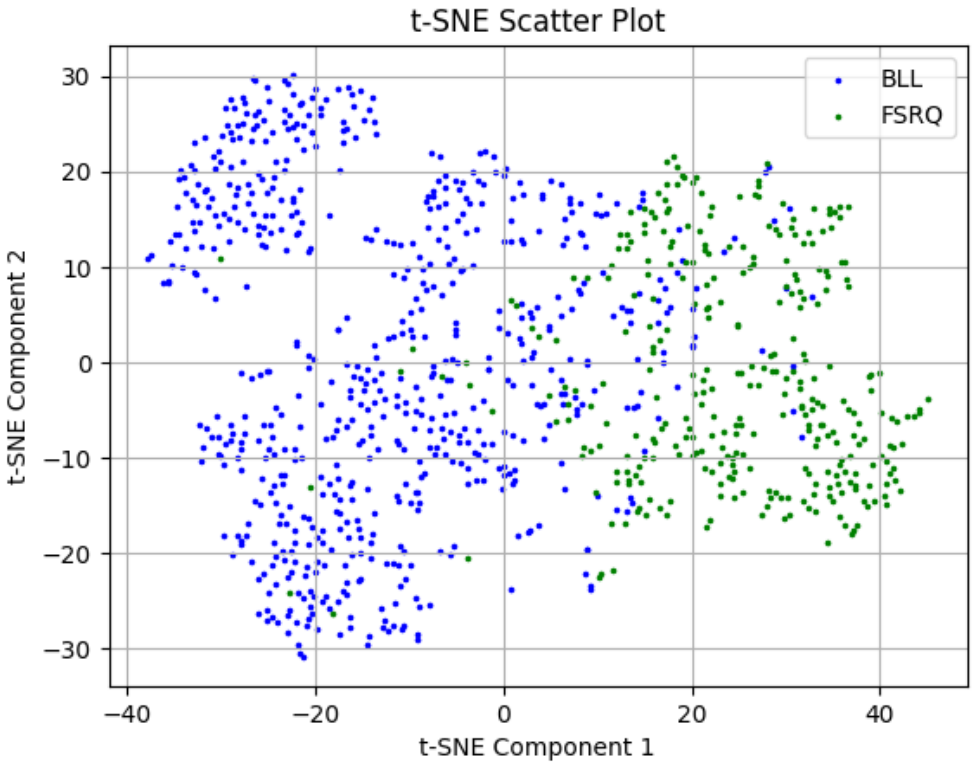


Figure 3.4: t-SNE scatter plot of multi-frequency analysis with labels from Fermi-LAT catalog

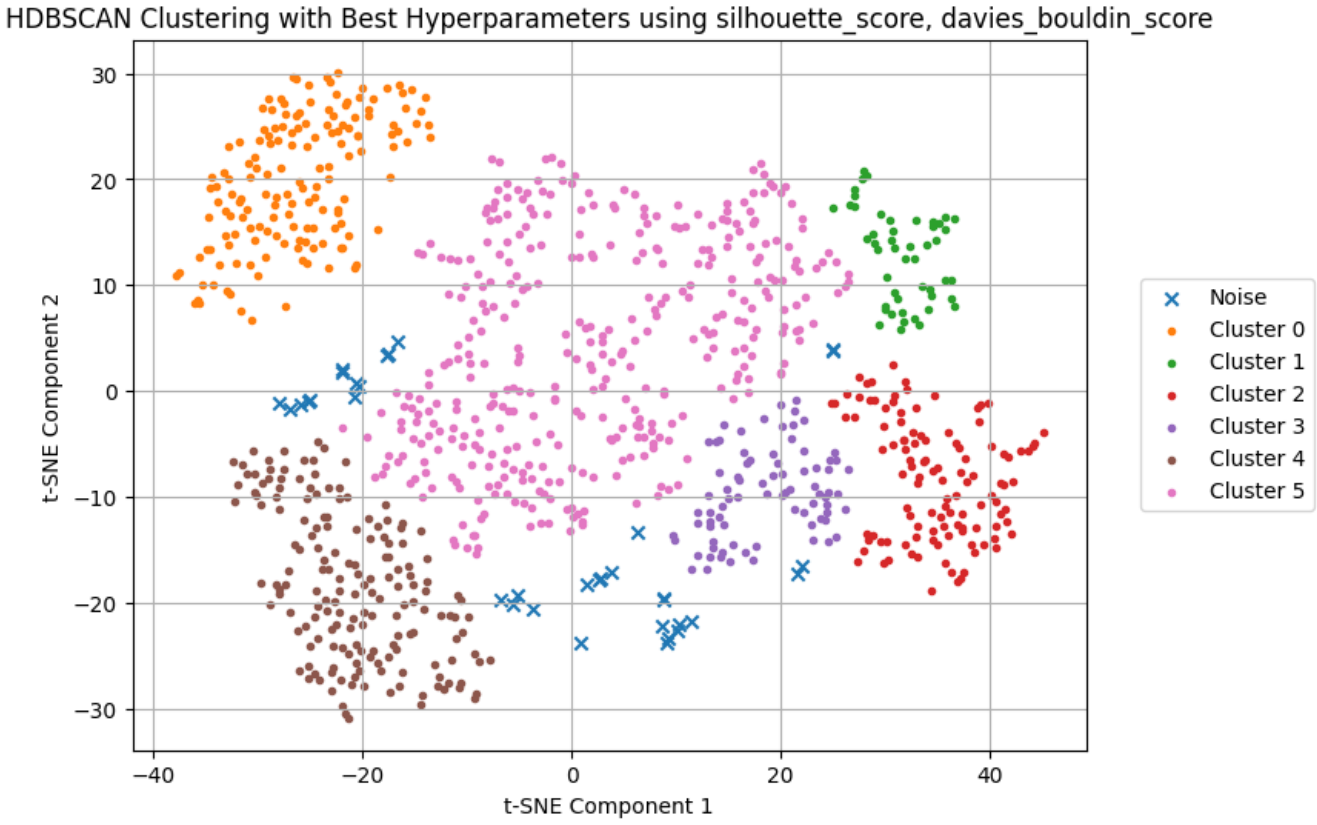


Figure 3.5: Clustering using hdbscan of a t-SNE scatter plot of multi-frequency analysis using silhouette_score, davies_bouldin_score. Best Hyperparameters: $min\_cluster\_size = 32$, $min\_samples = 1$, $cluster\_selection_epsilon = 0$. Best Silhouette $Score = 0.31$, Best Davies-Bouldin $Index = 1.235$

Hyperparameters used in HDBSCAN:

- **Min_cluster_size:** It is the minimum number of points required to form a cluster. Clusters with fewer points than this value will be labeled as noise or outliers. A larger min_cluster_size will result in fewer but more outlier-resistant and well-defined clusters, while a smaller value will allow smaller clusters to form, possibly capturing more fine-grained structures in the data. So 5 and 32 in the umap and t-SNE are the minimum number of points to form clusters.

- **Min_samples:** It is the minimum number of samples in a neighborhood for a point to be considered as a core point. Core points are crucial in the density-based clustering process, and they serve as seed points for cluster expansion. A higher value makes the algorithm more stringent in identifying core points, resulting in denser clusters with more points classified as noise.

- **Cluster_selection_epsilon:** It sets the cluster density threshold. Points within this distance of a cluster core point are considered part of that cluster. Setting it to zero makes the algorithm use the min_cluster_size to determine the cluster density, while a positive value allows for more relaxed density conditions, resulting in larger clusters.

Metric used for evaluation:

- **Silhouette_score:** It is used to evaluate the quality of clustering. It measures how similar an object is to its cluster compared to other clusters. It ranges from -1 to +1, where a higher value indicates that the object is well-clustered and distant from neighboring clusters, while a negative value suggests it might be misclassified. So I tried to increase it.

- **Davies_bouldin_score:** The Davies-Bouldin Index is another metric used for cluster evaluation. It measures the average similarity between each cluster and its most similar cluster, taking into account both cluster separation and compactness. Lower values indicate better-defined and well-separated clusters. So I tried decreasing it. Based on these two I chose the hyperparameters which gave the best result.

3.4 Supervised Machine Learning

Here in this section, we will use RandomForestClassifier, Support Vector Machine (SVM), Extreme Boost (XgBoost), and LightGBM are used here. This whole analysis is a binary classification problem between BLL and FSRQ. All are robust, do not require much of hyperparameter tuning and if we have multiple models we can get a better estimate of the unknown class (here it is BCU) via the ensemble method (explained below). Some key points are noted here:-

- **Class Weights:** Here we are trying to do binary classification but both classes don't have an equal number of members which creates a class imbalance. So the model will just assign most of them the class of the majority and since most of them belong to the majority so misclassification of the minority is a small price to pay for the majority getting correctly assigned. It will increase accuracy but this model is not generalized but clearly biased towards the majority. Hence assigning class weights i.e. higher weight to the minority class, so the model focuses on it more, and lower weight to the majority balances the model. It is done automatically using compute_class_weight from sklearn. It is also used in my analysis.

- **Ensemble method:** Here we will use various models and train them on the fully known dataset and then use it to predict for the unknown case. All the possible outcomes from each model and the number of iterations done for generalization. All these will be stored and based on hard voting i.e. assigning the class what the maximum suggests will be a more robust method. This is known as the ensemble method.

Here I did 5-fold cross-validation and also for 20 iterations for each method. k-fold cross-validation means that the overall dataset will be divided into k parts and k-1 parts will be used to train the model and the last part will be used to predict. And this rotation will be done over all parts. The prediction of each part is stored and we now have the full array to compare the results with the actual labels. This whole thing is achieved by defining the classifier and then giving the classifier, Data, Label, k-fold (number of folds), etc. to cross_val_predict. It will automatically do the mentioned steps by itself. StratifiedKFold is used to divide the whole data into k-folds almost equally to avoid bias. Both of these StratifiedKFold and cros_val_predict are part of sklearn.model_selection.

Here, the general behavior of each model for the given data is experimented with and realized. In the following sub-section, details related to the classifier (model) and the result are explained.

3.4.1 Random Forest

It is an ensemble of various decision trees each focusing on a different feature. We can control the number of trees, maximum depth, splitting after each node, etc. This is a robust method that is extensively used and it is not prone to over-fitting, unlike other methods. Each tree will predict and based on maximum voting the class will be decided.

Result:

Accuracy: 0.9351 ± 0.0032

Class	Precision	Recall	F1-score
BLL	0.9487 ± 0.0036	0.9542 ± 0.0037	0.9515 ± 0.0024
FSRQ	0.9074 ± 0.0069	0.8969 ± 0.0076	0.9021 ± 0.0048

Table 3.3: Classification Report for Random Forest 5-fold 20 iterations

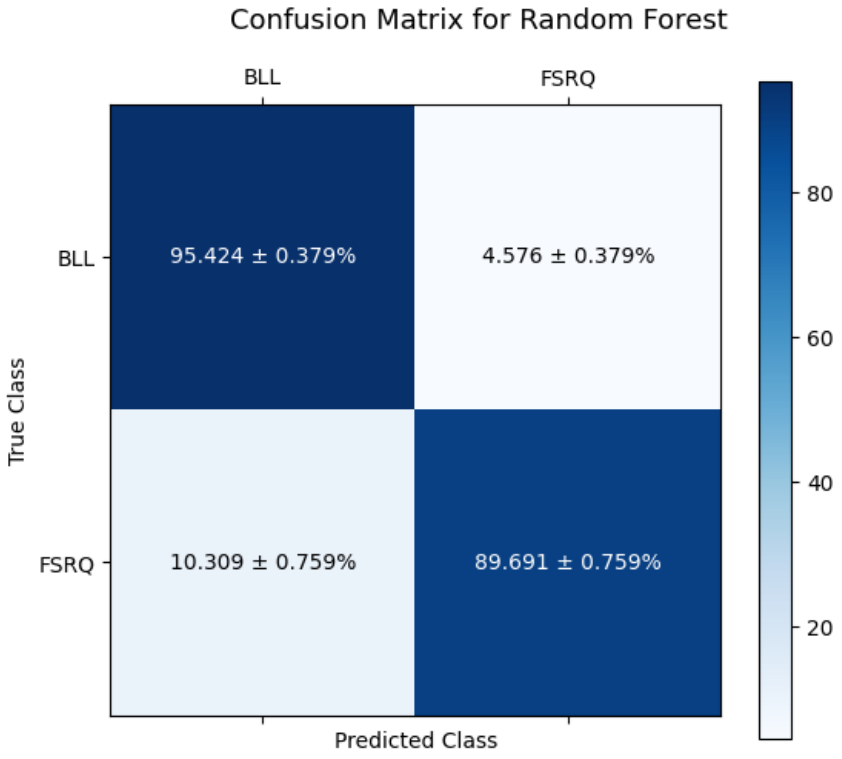


Figure 3.6:

3.4.2 Support Vector Machine (SVM)

It is a ML technique where using kernel function it implicitly projects the data to a higher dimension where linear separation is possible via a hyperplane. Its goal is to maximize the margin(distance between the hyperplane and the nearest data points (support vectors) from each class) but at the same time minimize the classification error. Regularization parameter C controls the trade-off between them. Support vectors are the data points that are closest to the decision boundary (hyperplane). Since it mainly depends on them hence it is a memory-efficient method.

Result:

Accuracy: 0.9269 ± 0.0026

Class	Precision	Recall	F1-score
BLL	0.9699 ± 0.0018	0.9188 ± 0.0039	0.9436 ± 0.002
FSRQ	0.8531 ± 0.006	0.9429 ± 0.0036	0.8958 ± 0.0034

Table 3.4: Classification Report for SVM 5-fold 20 iterations

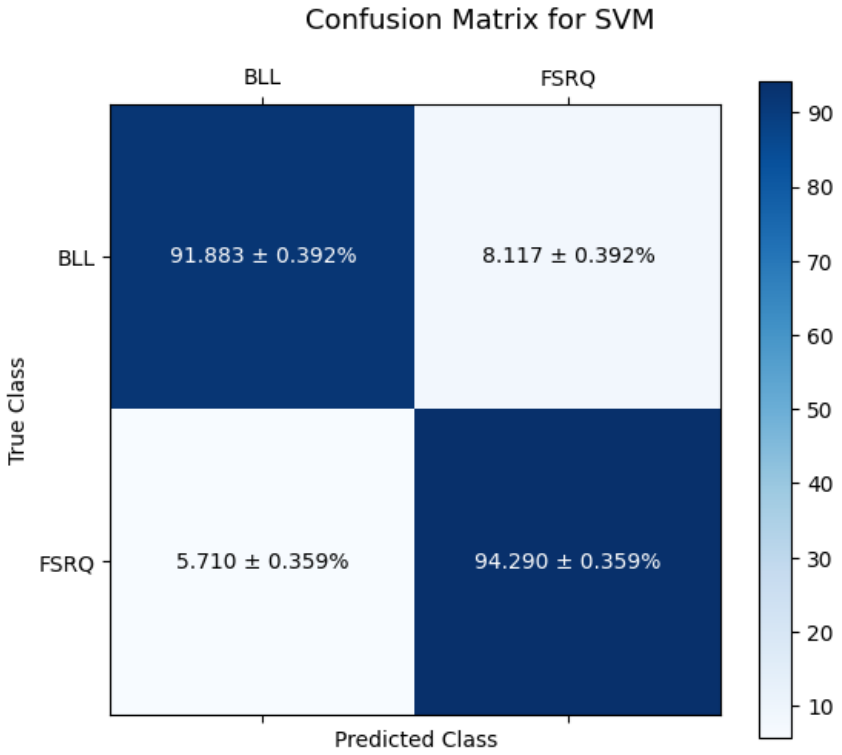


Figure 3.7:

3.4.3 XgBoost

XgBoost is an advanced gradient boost method where a lot of weak learners (decision trees) are built up one on another, correcting the errors made by the previous one. It uses gradient descent of first and second order for loss minimization and faster convergence. Hence it is one of the most popular methods.

Result:

Accuracy: 0.9321 ± 0.0043

Class	Precision	Recall	F1-score
BLL	0.9507 ± 0.0031	0.9474 ± 0.005	0.949 ± 0.0033
FSRQ	0.8956 ± 0.009	0.9017 ± 0.0064	0.8985 ± 0.0062

Table 3.5: Classification Report for XgBoost 5-fold 20 iterations

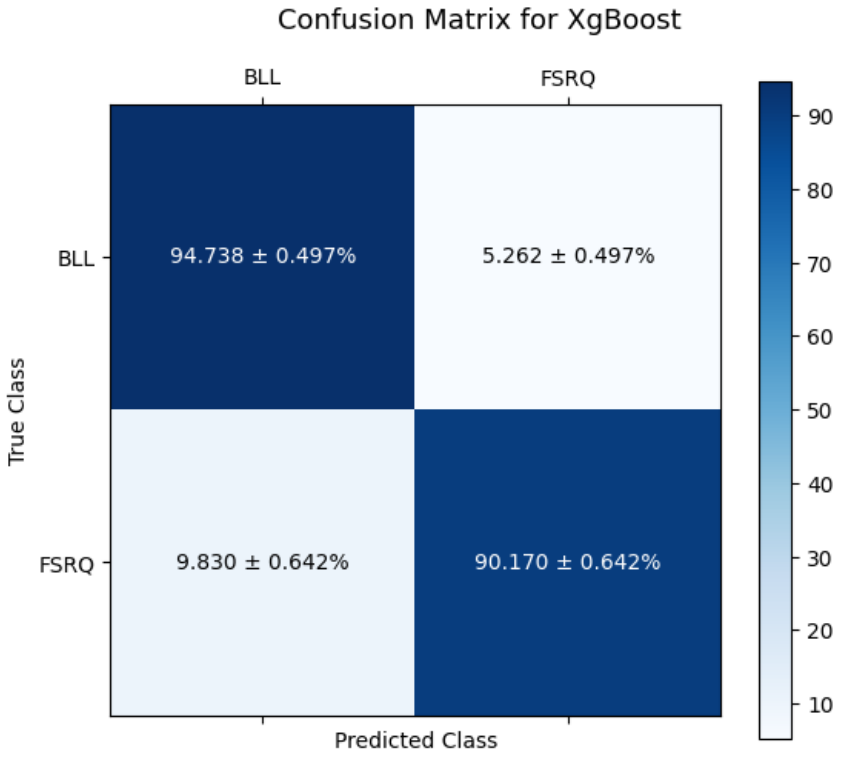


Figure 3.8:

3.4.4 LightGBM

LightGBM is a gradient boost method similar to XgBoost but the key difference is it grows leaf-wise rather than level-wise. It works by binning the data or histogram-based technique and it also selectively chooses points called "gradient-based one-sided sampling" which creates an instance that leads to faster convergence. Regularization, parallel computing, GPU acceleration,memory-efficient, fast etc are its key features because of which it is really popular.

Result:

Accuracy: 0.9374 ± 0.0045

Class	Precision	Recall	F1-score
BLL	0.9541 ± 0.0039	0.9519 ± 0.0041	0.953 ± 0.0034
FSRQ	0.9043 ± 0.0077	0.9084 ± 0.008	0.9064 ± 0.0067

Table 3.6: Classification Report for LightGBM 5-fold 25 iterations

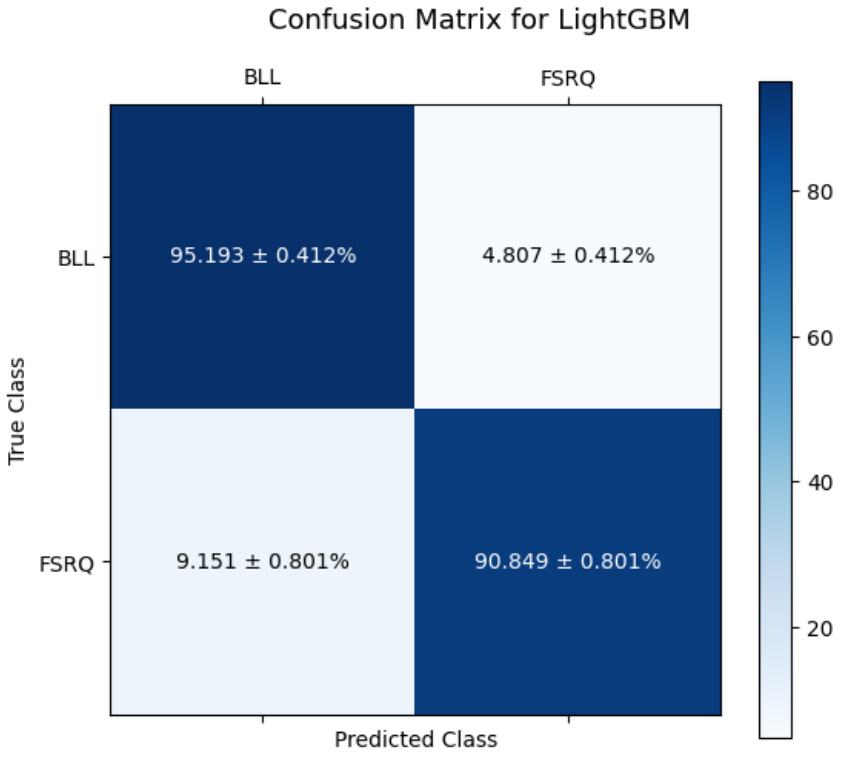


Figure 3.9:

Also, let's see the ROC curve. Now, to further analyse the importance of each column we will plot the "relative feature importance". Since it is can only be done for tree-based methods, hence we are excluding SVM from this analysis.

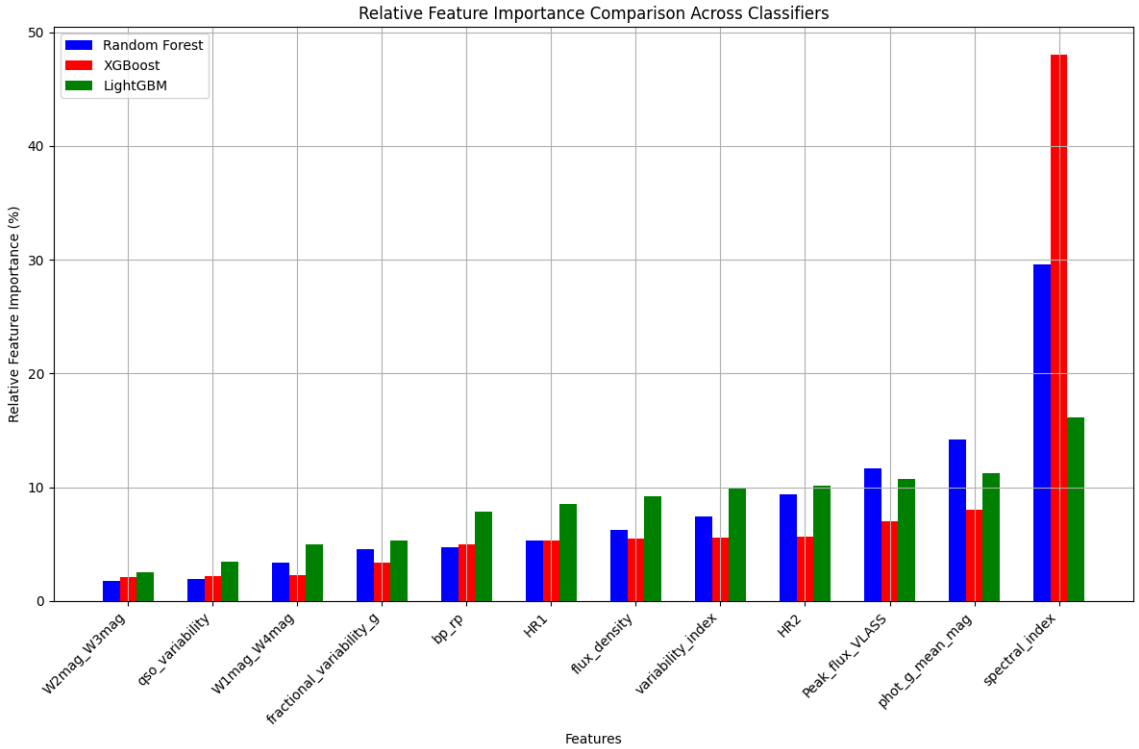


Figure 3.10:

Based on this we can conclude that the spectral index from Fermi-LAT is the key feature in this classification. And the peak flux from VLASS also in the top 3 feature indicating that the VLA data is indeed useful in the classification of blazars using machine learning.

Chapter 4

Epoch 1 and 2 comparison

4.1 Introduction

There are two types of transients: variables and one-time transients.

- **Variable transients:** They are those events in which stellar objects undergo changes in brightness over time and their study can reveal their intrinsic properties. It includes various types of stars, such as Cepheid variables, RR Lyrae stars, and eclipsing binary stars, quasars, etc.
- **One-time transients:** These events occur only once and it is not possible to see them again if not observed at the time of their occurrence. Examples are gamma-ray bursts, fast radio bursts, supernovae, etc.

Analysis between two epochs tells us a lot about transients. There are three possibilities a source is in epoch 1 and not in epoch 2, a source in epoch 2 but not in epoch 1 (either not present or went below detection level), and those that are there in both but the difference between their flux above a threshold. All these cases will be discussed in this chapter. This work is still in its initial phase and basic work done is covered here.

4.2 Sources belonging to epoch one only

By taking epochs 1 and 2 and removing all counterparts (one or more within 2.5 arcsecs).After that the constraint that $peak\_flux > 3$ and $deconvolved_major < 10$arcsecs is applied. As a result, I got 26,308 sources. Out of which 10,040 have NED counterparts. The full table of types is as follows:

Table 4.1: Description of Various Astronomical Objects from NED counterparts for constraint epoch 1 only

Element	Count	Description
Not Available (nan)	16268	Data point with no NED counterpart
Irregular Variable Star (IrS)	2811	Irregularly varying stars
Galaxy (G)	1252	Galaxies observed in the survey
UV-Excess Star (UvS)	295	Stars with excess ultraviolet emission
Radio Source (RadioS)	5146	Celestial objects emitting radio waves

Continued on next page

Table 4.1 – Continued from previous page

Element	Count	Description
Quasar (QSO)	165	Extremely bright and distant active galactic nuclei
UV-Excess Star in the Far Ultraviolet (UvES)	14	Stars with excess far-ultraviolet emission
Galaxy Cluster (GClstr)	78	Group of galaxies held together by gravity
X-ray Source (XrayS)	51	Celestial objects emitting X-rays
Blue Star (Blue*)	1	Bright blue star
Non-Stellar Source (!*)	4	Non-stellar objects detected in the survey
Absorption Line System (AbLS)	3	Systems showing absorption lines in their spectra
Visible Source (VisS)	33	Objects visible in the survey
Star (!*)	171	Stars detected in the survey
Galaxy Pair (GPair)	1	Pair of galaxies in close proximity
Star Cluster (*Cl)	2	Cluster of stars
Galaxy Triple (GTrpl)	1	Set of three galaxies
Planetary Nebula (PN)	2	Luminous shells of gas and dust ejected by dying stars
HII Region (HII)	4	Region of ionized gas around hot stars
Other	1	Other types of objects not categorized above
Supernova (SN)	3	Exploding stars
Non-Variable Star (!V*)	1	Non-variable star
Non-Variable Planetary Nebula (!PN)	1	Non-variable planetary nebula

Here as we can see that HII , SN, PN are there which should not show such behaviour and thus their analysis is done using VLASS cutout of that region of 1 arcmin square around the given coordinate. They are mentioned in the table below for epoch 1 only.

4.3 Sources belonging to epoch 2 only

Similar to epoch 1, taking epochs 1 and 2 and removing all counterparts (one or more within 2.5 arcsecs).After that the constraint that $peak\_flux > 3$ and $deconvolved_major < 10$arcsecs is applied. As a result, I got 48,521 sources. Out of which 14,237 have NED counterparts. The full table of types is as follows:

Table 4.2: Description of Various Astronomical Objects from NED counterparts for constraint epoch 2 only

Element	Count	Description
Not Available (nan)	34284	Data point with no NED counterpart
Irregular Variable Star (IrS)	4083	Irregularly varying stars
Radio Source (RadioS)	6812	Celestial objects emitting radio waves
Galaxy (G)	1871	Galaxies observed in the survey
Star (!*)	343	Stars detected in the survey
Galaxy Cluster (GClstr)	135	Group of galaxies held together by gravity
Quasar (QSO)	300	Extremely bright and distant active galactic nuclei
UV-Excess Star (UvS)	473	Stars with excess ultraviolet emission
Visible Source (VisS)	61	Objects visible in the survey
X-ray Source (XrayS)	78	Celestial objects emitting X-rays
UV-Excess Star in the Far Ultraviolet (UvES)	22	Stars with excess far-ultraviolet emission
Variable Star (V*)	1	Variable star
Planetary Nebula (PN)	9	Luminous shells of gas and dust ejected by dying stars
Non-Variable Star (!V*)	7	Non-variable star
HII Region (HII)	7	Region of ionized gas around hot stars
Other	1	Other types of objects not categorized above
Group of Galaxies (GGroup)	7	Group of galaxies
Supernova (SN)	6	Exploding stars
Star Cluster (*Cl)	7	Cluster of stars
Absorption Line System (AbLS)	6	Systems showing absorption lines in their spectra
Non-Stellar Source (!*)	3	Non-stellar objects detected in the survey
Point of Light in a Galaxy (PofG)	1	Point-like source within a galaxy
Galaxy Pair (GPair)	3	Pair of galaxies in close proximity
Pulsar (Psr)	1	Pulsating neutron star

Here as we can see that HII , SN, PN are there which should not show such behaviour and thus their analysis is done using VLASS cutout of that region of 1 arcmin square around the given coordinate. They are mentioned in the table below for epoch 2 only.

4.4 Analysis using DS9

Here, the analysis of HII, SN, and PN is done via VLA cutout images because these should not show transient behavior. Hence the image around the coordinate with the size of (1 arcmin) X (1 arcmin) image. And corresponding images from all 3 epochs (3rd epoch data is not out yet but it is somehow available there). Now using DS9 I loaded images parallelly and smoothed them, made a grid, did frame matching, made contour, and selected color scale such it would be easy to analyze. Then I saved those images for future analysis. Also, the shape, and center of the bright region's coordinates are also listed down. And comparative distances are also noted down.

Key takeaway points are:

- Most of them had a counterpart in the other epoch as only that single source was visible even if it was much greater than 2.5 arcsec (the cutout was of 1 arcmin). In case multiple points are there I have considered them too.
- Their distinct features such as lobes or a sphere etc are visible.
- If something is visible in both epochs then we can clearly observe the size and brightness reducing in the next epoch.
- There were a few cases where either the image in epoch 1 or 2 is too noisy or if it is only in epoch 1 or 2 then there is an observation in epoch 3. So we can check that why it was like that.
- Even though data of epoch 3 is not there in public, it is in the VLASS cutout.

Chapter 5

Future Work

Further investigation is needed to reconcile the numerical values in Chapter 2 with the detailed data even though in the analysis the numbers were self-consistent. Unsupervised learning analysis yielded inconclusive results, prompting the need for additional studies. Spectral Energy Distribution (SED) analysis could provide further insights. A comprehensive analysis of the transients in Chapter 4 is recommended, as this work is still in its initial stages. When epoch 3 data becomes available, a more detailed and thorough analysis can be conducted.

Chapter 6

Conclusion

The work I did in this internship taught me a lot of things. Not only certain skills but it made me understand how research work is done since this project was mainly exploratory in nature I tried a lot of other things many are there in this report and few are not. This made me think about all the possible approaches I can take and things that can be done on top of what my guide has suggested to me. Making a report in latex, time to time presentations improved my presentation skills, note-making to keep track of my discussion with my mentor, etc. Overall the internship fulfilled its role by giving me exposure in the field of astroinformatics.

The technical skills I learned during this project are big data handling, knowledge of various astronomical catalogs and how to do cross-catalog analysis, plotting, handling of GPU cluster, unsupervised and supervised machine learning, TOPCAT and Python.

Acknowledgment for Catalogs

This research has made use of the CIRADA cutout service at URL cutouts.cirada.ca, operated by the Canadian Initiative for Radio Astronomy Data Analysis (CIRADA). CIRADA is funded by a grant from the Canada Foundation for Innovation 2017 Innovation Fund (Project 35999), as well as by the Provinces of Ontario, British Columbia, Alberta, Manitoba and Quebec, in collaboration with the National Research Council of Canada, the US National Radio Astronomy Observatory and Australia's Commonwealth Scientific and Industrial Research Organisation.

This research has made use of NRAO FIRST "The National Radio Astronomy Observatory is a facility of the National Science Foundation operated under cooperative agreement by Associated Universities, Inc."

This research has made use of Fermi LAT.The Fermi LAT Collaboration acknowledges generous support from a number of agencies and institutions that have supported the Fermi LAT project. These include the National Aeronautics and Space Administration and the Department of Energy in the United States; the Commissariat à l'Energie Atomique and the Centre National de la Recherche Scientifique/Institut National de Physique Nucléaire et de Physique des Particules in France; the Agenzia Spaziale Italiana and the Istituto Nazionale di Fisica Nucleare in Italy; the Ministry of Education, Culture, Sports, Science and Technology (MEXT) and the Japan Aerospace Exploration Agency (JAXA) in Japan; and the K. A. Wallenberg Foundation, the Swedish Research Council and the National Space Board in Sweden.

This research has made use of VLASS. The Very Large Array Sky Survey (VLASS) is a project of the National Radio Astronomy Observatory (NRAO). NRAO is a facility of the National Science Foundation operated under a cooperative agreement by Associated Universities, Inc.

This research has made use of the 2SXPS catalog. The 2SXPS catalog is based on data obtained from the ROSAT All-Sky Survey (RASS), which was made possible by the support of the German Bundesministerium für Bildung und Forschung (BMBF) and the Max-Planck-Gesellschaft (MPG).

This research has made use of the SDSS catalog. Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England.

This research has made use of the NED catalog. The NASA/IPAC Extragalactic Database (NED) is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

This research has made use of the GAIA catalog. The Gaia mission is a project of the European Space Agency (ESA). The data processing and archive operations are mainly carried out by the Gaia Data Processing and Analysis Consortium (DPAC), a collaboration of about 400 scientists and engineers in 20 institutes across Europe.

This research has made use of the WISE catalog. The WISE mission is a project of the Jet Propulsion Laboratory (JPL), California Institute of Technology, funded by the National Aeronautics and Space Administration (NASA).

References

[1] Gordon, Y. A., Boyce, M. M., O'Dea, C. P., Rudnick, L., Andernach, H., Vantyghem, A. N., Baum, S. A., Bui, J.-P., Dionyssiou, M., Safi-Harb, S., Sander, I. (2021). A Quick Look at the 3 GHz Radio Sky. I. Source Statistics from the Very Large Array Sky Survey. The Astrophysical Journal Supplement Series, 255, 30. https://doi.org/10.3847/1538-4365/ac05c0

[2] https://umap-learn.readthedocs.io/en/latest/how_umap_works.html

[3] https://www.ibm.com/topics/machine-learning

[4] https://www.star.bris.ac.uk/ mbt/topcat/