Using radio (VLASS) data for better classification of AGNs in Fermi LAT 14-year catalog

SATYAPRIYA DAS[1,*] ,RESMI LEKSHMI[1,*]

[1] INDIAN INSTITUTE OF SPACE SCIENCE AND TECHNOLOGY, TRIVANDRUM – 556947

* Email: satyapriya1203@gmail.com, l.resmi@gmail.com

Abstract

The Fermi Large Area Telescope (LAT) has detected more than 6000 gamma-ray sources in the first 14 years of its operation. The major fraction of LAT point sources are Active Galactic Nuclei (AGN) and pulsars. Blazars constitute the largest portion of Fermi AGN and are divided into BL Lacertae objects (BL Lac), Flat Spectrum Radio Quasars (FSRQs), and blazars of unidentified type (BCUs). For a robust classification, one needs to consider not only the gamma-ray data but also the multi-wavelength information. In past, Swift, Gaia, SDSS, and WISE counterpart data are used for Fermi AGN classification. After the release of the VLA Sky Survey (VLASS) data, it is possible to supply information of radio counterparts also into the classification algorithms. We use machine learning techniques to classify AGN in the 4th data release of the Fermi LAT point catalogue (4FGL-DR4), supported by potential multi-wavelength counterparts from Swift BAT/XRT, Gaia, SDSS, WISE, and VLASS.

Background

Introduction:

- Blazars are among the most enigmatic and energetic objects in the universe, they are a 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 analysis.

Motivation:-

Blazars emit radiation across a broad range of frequencies, from radio to gamma-ray. ML algorithms are adept at handling multifrequency data and extracting complex patterns from diverse datasets, enabling a comprehensive analysis of blazar emissions.

Software and Catalogs

Software: TOPCAT

Language: Python

Packages: Astropy, pandas, NumPy, Matplotlib

Catalogs: 4th Fermi LAC high galactic latitude **[2]**, GLEAN **[3]**(GAIA AGN catalog), 2SXPS catalog **[4]**, VLASS **[1]**, VLBI source position catalog.

Cross Catalog Matching

The coordinates of the counterpart catalog mentioned in Fermi LAC are used to ellipse match with various other catalogs using quadruple match in TOPCAT. The VLBI source position catalog is matched using the mentioned VLBI name in the Fermi LAC catalog itself.

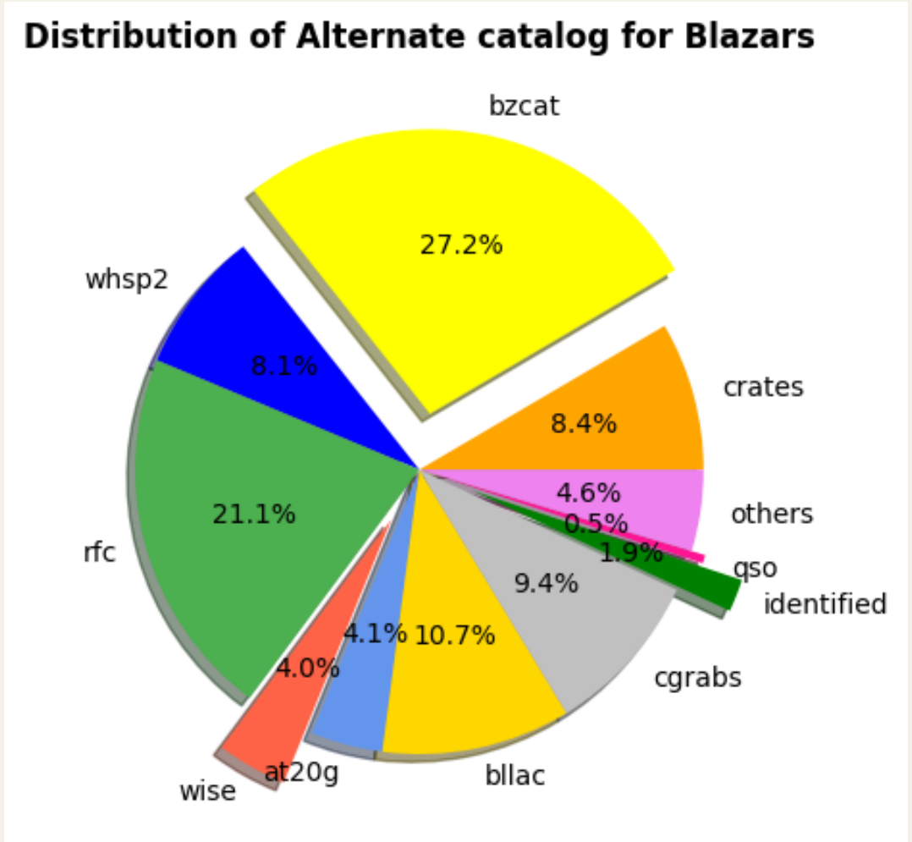

Fig1: Pie chart depicting various catalogs used in alternate catalog in FermiLAC

Data Cleaning & Preprocessing

Columns with photometric data are selected for the analysis. It is cleaned and their correlation map is plotted to remove the highly correlated columns and only keep one of them based on its importance and availability out of all rows. The finalized columns are shown in Table 1. Column with * are kept and analysis is done with and without them.

After that the following constraints are applied:-
significance>4 and error_radius_counterpart < 5 arcsec (Fig 2)

Various imputation methods are compared and median imputation gave the best result. Hence, that is used in the subsequent parts. For the normalization, values are scaled between 0 to 1.

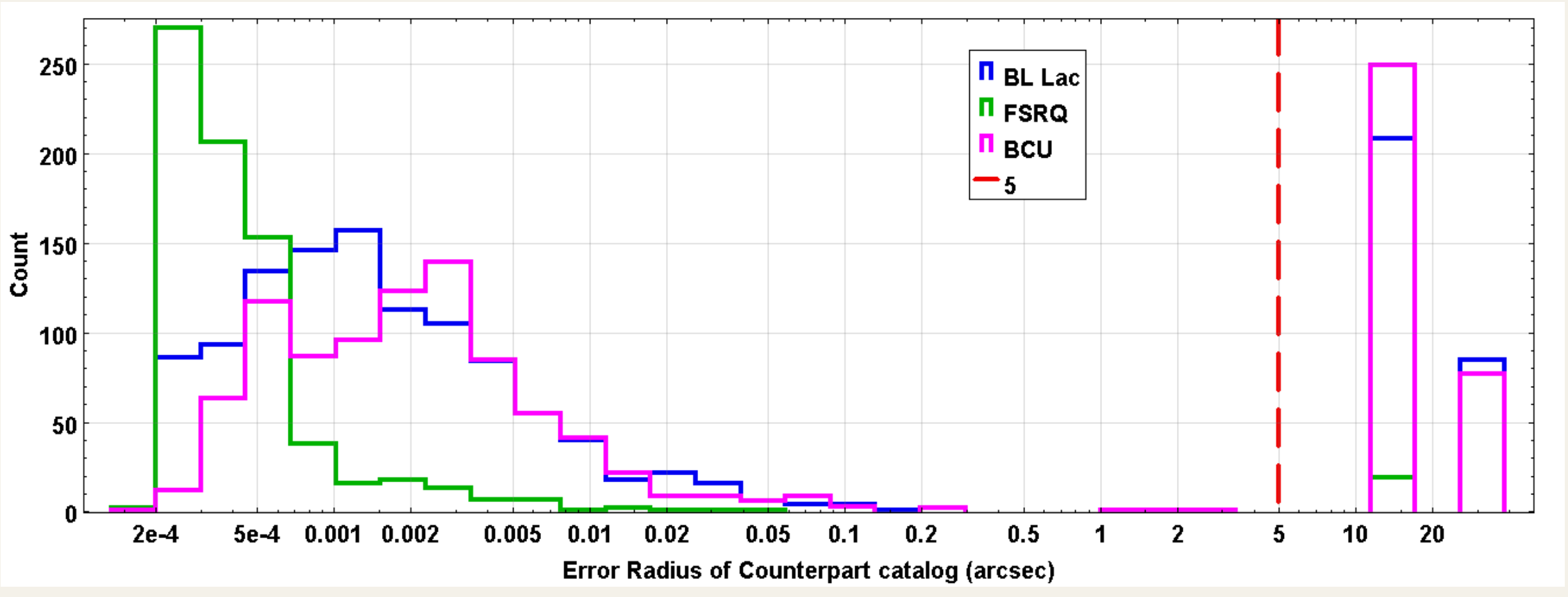

Fig 2: Distribution of error of counterpart catalog

Final Data

Table 1: Columns selected for Machine Learning

Column Name	Count	Catalog	Description
Energy_Flux100	2668	FermiLAC	Energy flux from 100 MeV to 100 GeV obtained by spectral fitting.
Fermi_index	2668		Combined index; by selecting the index from various other model indexes depending on its spectral type.
Pivot_Energy	2668		The energy at which the error on differential photon flux is minimal (decorrelation energy for the PowerLaw fit).
HE_nuFnuPeak*	2281		The quantity nu * fnu at the high-energy-peak frequency, in erg/cm2/s.
nu_syn	2668		The synchrotron-peak frequency in the observer frame, in Hz.
Variability_Index	2668		The sum of the log(likelihood) difference between the flux fitted in each time interval and the average flux over the full catalog interval.
Frac_Variability	2668		The fractional variability computed from the fluxes in each year.
X-band Total	2486	VLBI Source Position Catalog	Flux in X-band from VLBI Source Position Catalog.
phot_g_mean_mag	1662	GAIA AGN Catalog (GLEAN)	G-band mean magnitude from GAIA.
bp_rp	1662		Magnitude in Gaia Bp—Magnitude in Gaia Rp.
PowFlux	1280	2SXPS Catalog	The observed 0.3—10 keV flux (in erg cm-2 s-1) derived from the power-law spectrum.
InterpPowGamma	807		The power-law photon index of the spectrum of the source derived from the hardness ratios of the source interpolated on a look-up table of power-law spectra.
Total_flux_VLASS*	1984	VLA Sky Survey	Total flux from VLA Sky Survey at 3 GHz.
Class	2668	FermiLAC	Blazar classes - BLL : BL Lac , FSRQ: Flat Spectrum Radio Quasar, BCU: Blazar Candidate of Unknown Type (will be classified into above two).

Machine learning Analysis

Tree-based classifiers namely Random Forest, XgBoost, and LightGBM are used for this analysis. They are selected because they are robust to outliers and can find complex patterns without even the need for imputation. Random forest is made of decision trees, each tree draws samples with replacements from various rows and columns, gives a prediction, and based on voting the final classification is made. XgBoost grows trees sequentially and updates based on previous mistakes. LightGBM grows leafwise, choosing those with maximum delta loss.

The effect of with and without imputation on Useful & Full (Table 2 description), and with & without radio data (total 8 cases) after optimization for all 3 models is shown below in Table 2. In all cases, they are normalized, and values are scaled between 0 to 1.

Table 2: Train and test accuracy in various cases. Useful are columns without * columns. Full includes all.

		Random Forest		XgBoost		LightGBM	
Method	**Removed col**	**Cross-val Accuracy**	**Test Accuracy**	**Cross-val Accuracy**	**Test Accuracy**	**Cross-val Accuracy**	**Test Accuracy**
Useful	-	97.26	89.26	96.58	88.15	96.94	84.85
Useful	X-band Total	96.89	91.18	95.7	87.88	96.26	83.2
Useful Imputed	-	97.27	90.08	96.24	87.05	96.96	88.98
Useful Imputed	X-band Total	96.73	88.71	96.16	90.63	96.26	88.98
Full	-	97.09	88.43	96.58	88.98	97.17	86.23
Full	X-band Total, VLASS	96.81	89.53	96.15	89.53	96.77	82.92
Full imputed	-	97.36	88.43	97.02	88.98	97.09	89.81
Full imputed	X-band Total, VLASS	96.82	90.63	96.74	90.63	96.26	88.98

AutoML from Flaml is used to optimize each model in all cases. The data is divided into train (80%) and test (unseen data, 20%). Cross-validation is done on train data and the configuration used is as follows:-

- Time budget = 100 sec
- Class weights are applied, task = classification, metric = roc_auc
- Cross-validation = 5 fold via Stratified KFold (equal distribution of both classes in each fold)

ROC AUC score shows how well the classifier distinguishes positive and negative classes. It can take values from 0 to 1. A higher ROC AUC indicates better performance. A perfect model would have an AUC of 1, while a random model would have an AUC of 0.5. Here roc_auc = 1- roc_auc_score. This means maximizing the shaded area by minimizing the unshaded part.

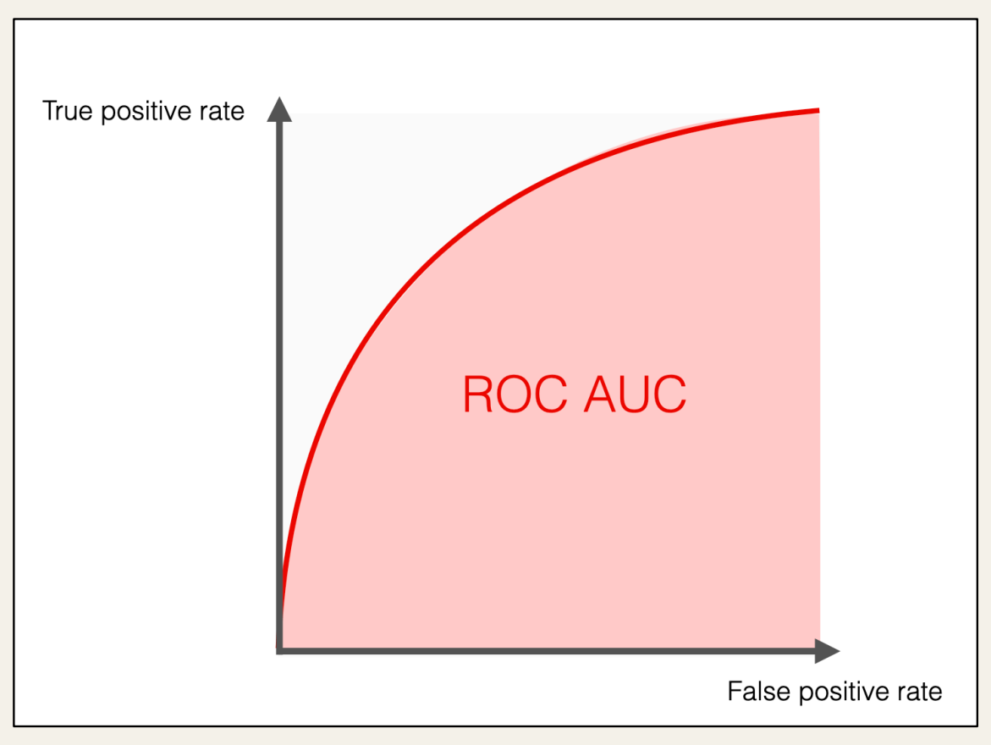

Fig 3: ROC AUC schematic

In each case, permutation importance and feature importance are also calculated and ranked in descending order of importance.

- **Feature Importance:** X-band Total flux is 3rd to 6th and Total_Flux_VLASS is 7th to 10th rank.
- **Permutation importance:** X-band Total flux is around 5th to 8th most significant and Total_Flux_VLASS is 7th to 12th rank.

Acknowledgement

I acknowledge the Indian Institute of Space Science and Technology, Trivandrum for providing me with the opportunity. I appreciate all the peers and professors in my department for their invaluable discussions.

References

[1] Yjan A. Gordon *et al* 2021 *ApJS* **255** 30
[2] M. Ajello *et al* 2020 *ApJ* **892** 105
[3] Carnerero, Maria I., et al. *A&A* 674 (2023): A24.
[4] Evans et al., 2020, ApJS, 247, 54

Conclusion

- Supervised ML gave 97% accuracy on cross-validation and 89% on test dataset.
- Fermi spectral index, pivot energy, nu_syn, bp_rp, and X-band Total are most important for this classification.
- X-band data from VLBI is much more significant than Total_Flux_VLASS here.