ICRANet-Armenia

MACHINE LEARNING BASED CLASSIFICATION OF BLAZAR CANDIDATES OF UNCERTAIN TYPE(BCUs)

V. Vardanyyan M. Khachatryan N. Sahakyan

Introduction

Blazars

Blazars are radio-loud active galactic nuclei (AGNs) with a relativistic jet pointing towards the observer. These sources are divided into two main classes: BL Lacertae objects (BL Lacs) and Flat Spectrum Radio Quasars (FSRQs), which show very different optical spectra.

Fermi Large Area Telescope

The Fermi Large Area Telescope (LAT) has been continuously observing the γ -ray sky since 2008 August in the 100 MeV– 300 GeV energy range. The latest Fermi-LAT catalog is the LAT 10-yr Source Catalogue 4FGL-DR2, which lists 5788 γ-ray sources on four years of data. Out of the 5788 4FGL sources, 3437 are blazars: 1190 BL Lacs, 730 FSRQs, and 1517 blazar candidates of uncertain type (BCUs)

OBJECTIVE

Classifying BCUs remains a strategic goal

When optical spectra or multiwavelength information needed for a rigorous classification are not available, a statistical approach to the problem, including **machine learning**, can be very useful for the classification of BCUs

What is Machine Learning?

Machine learning is a method of recognizing patterns within data in order to achieve goals such as classification. In a type of machine learning called supervised machine learning, an algorithm classifies unknown objects by comparing their characteristics with characteristics of known objects.

Basically, ANN is a mathematical function over an Ndimensional space, where N is the number of input parameters to the network. Input parameters are values which describe an object (blazars in our case). ANN produces a likelihood for the object to belong to a certain class (when ANN is used for classification). The network is trained on already classified objects (known BL Lacs and FSRQs in our case)

Data used for classification

For input parameters, we used γ -ray light curves and spectra present in the 4FGL catalog.

10 time-integrated fluxes corresponding to 1-yr observation periods sorted by increasing value

energy integrated flux values in 6 different energy bands.

This produced a set of N = 16 input parameters to the network for each source

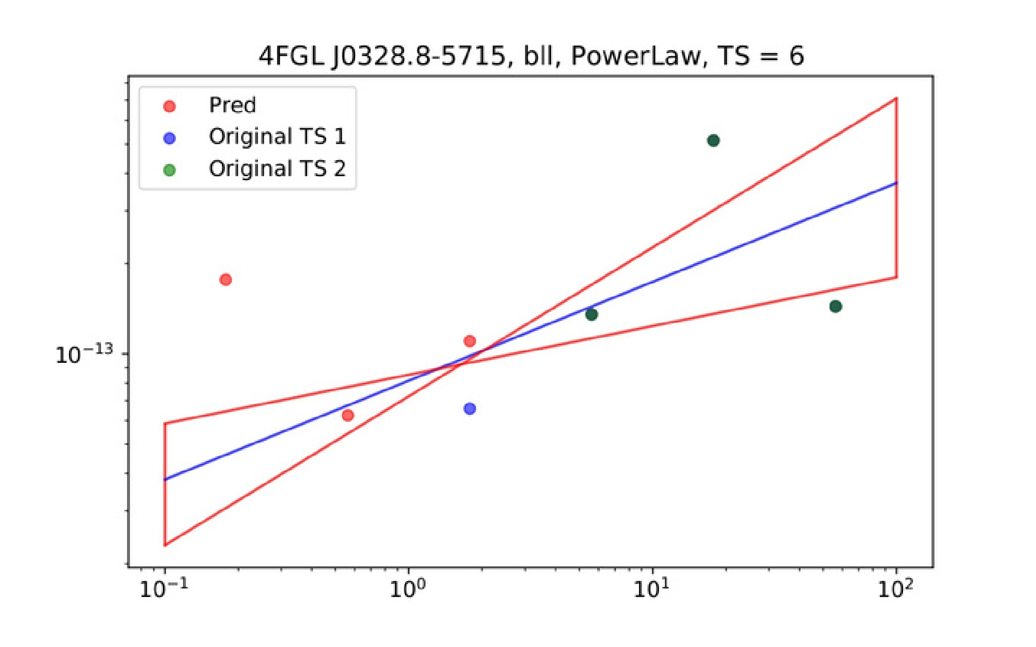
Spectral Energy Distribution

The 4FGL catalogue contains time-integrated fluxes in seven energy bands: 0.05–0.1, 0.1–0.3, 0.3–1, 1–3, 3–10, 10– 30, 30–300 GeV. We used six of them, starting from the second.

A quick comparison between BL Lacs and FSRQs shows several features: there is a difference in slope, i.e. average power-law index, with BL Lacs having a lower one;
BL Lacs on average have higher flux values than FSRQs in the highest energy band and vice versa for lowest;

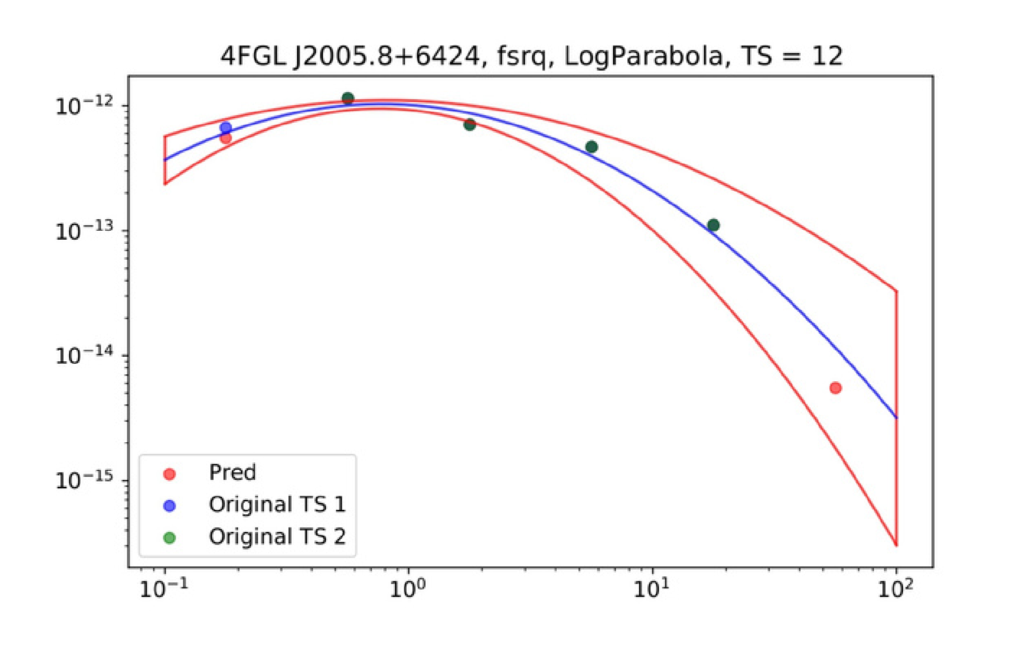
Among the considered sources some of them have missing data that is low test statistics in one of the measured energy intervals. In order to build a strong network initially, we fill the missing data points using a machine learning algorithm.

Examples of filled data



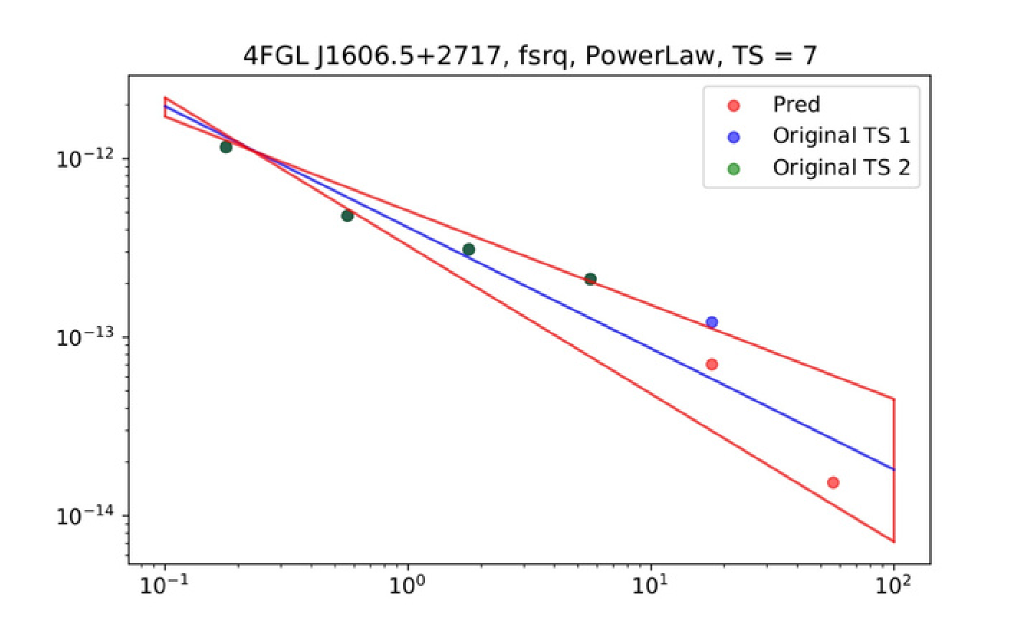
Example 1

Prediction (red dot) for PowerLow SED with BLLac source type compared to appeared point on low statistics (blue dot)



Example 2

Prediction (red dot) for LogParabola SED with FSRQ source type compared to appeared point on low statistics (blue dot)



Example 3

Prediction (red dot) for PowerLow SED with FSRQ source type compared to appeared point on low statistics (blue dot)

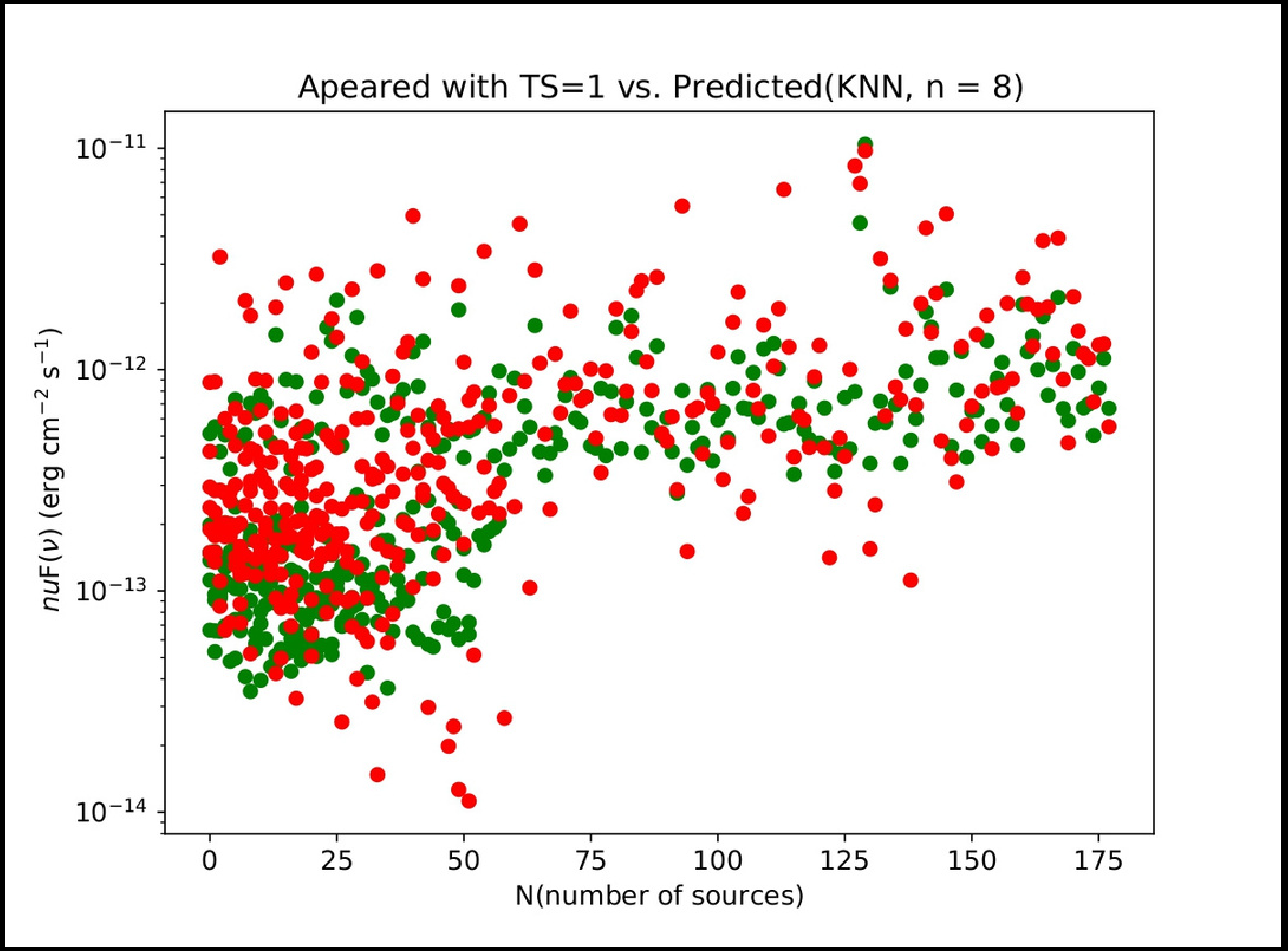
Fill Missing SED data

KNN imputation

The idea in kNN methods is to identify 'k' samples in the dataset that are similar or close in the space. Then we use these 'k' samples to estimate the value of the missing data points. Each sample's missing values are imputed using the mean value of the 'k'-neighbors found in the dataset

Also tried more advanced methods like MICE, MIDA, GINN, but as we have a limited dataset of around 1900 blazars with known source types, more complicated missing data imputation algorithms stay underfitted.

After implementing the knn method and comparing it to values from low test statistics from 1 to 2 TS range, we can see that our method can mimic the pattern of real data.



Light Curves

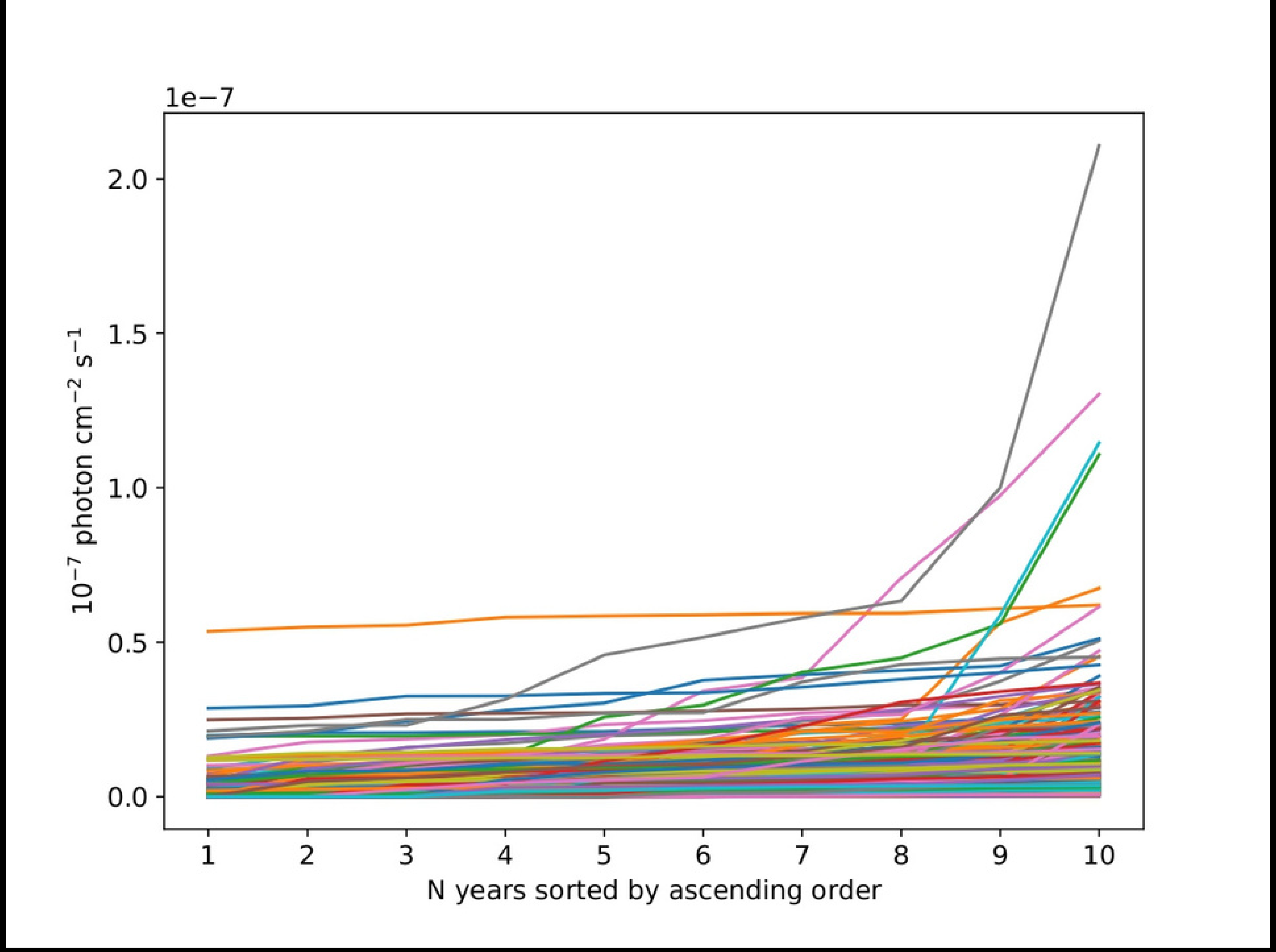
The 4FGL catalog contains light curves with a bin duration of 1 yr. This created a set of (1 yr × 10 yr) ten energy integrated fluxes for each source

Sorting the flux values from lowest to highest is one way of making blazar activities comparable. The ten annual time bins corresponding to 10 yr of Fermi-LAT observations are random time intervals in the life of each blazar.

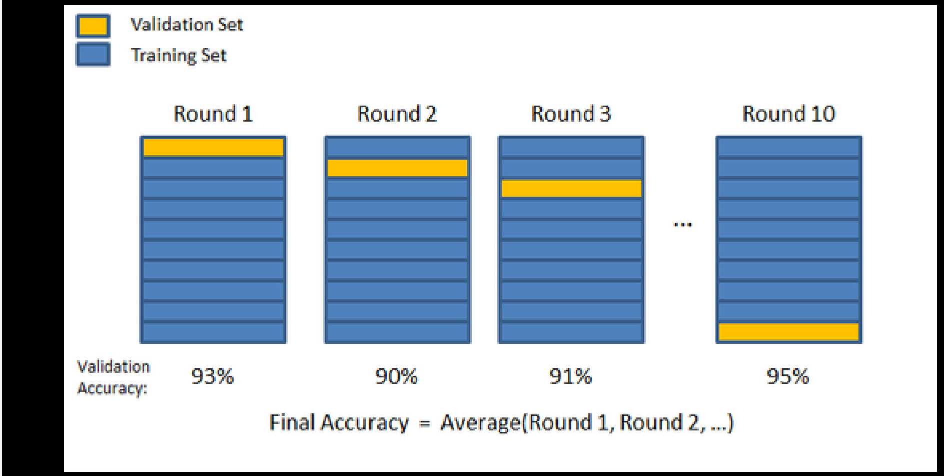
BL Lacs are on average dimmer than FSRQs in the Fermi-LAT energy range. Their activity tends to be more continuous over time than that of FSRQs.

Light curve

By sorting the flux values, we are directly comparing fluxes of dimmest, average, and brightest periods for each blazar and relationships between them.



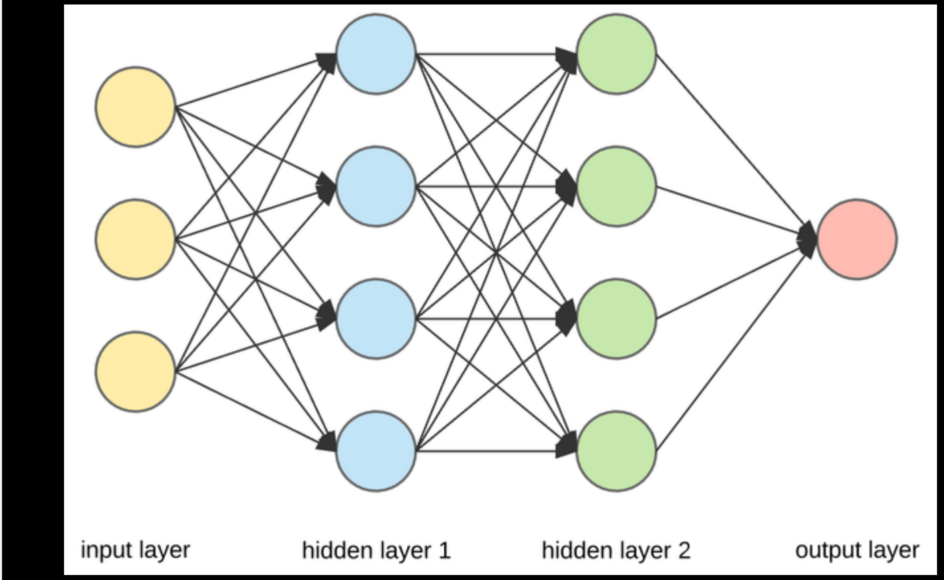
Model Selection



Cross Validation

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

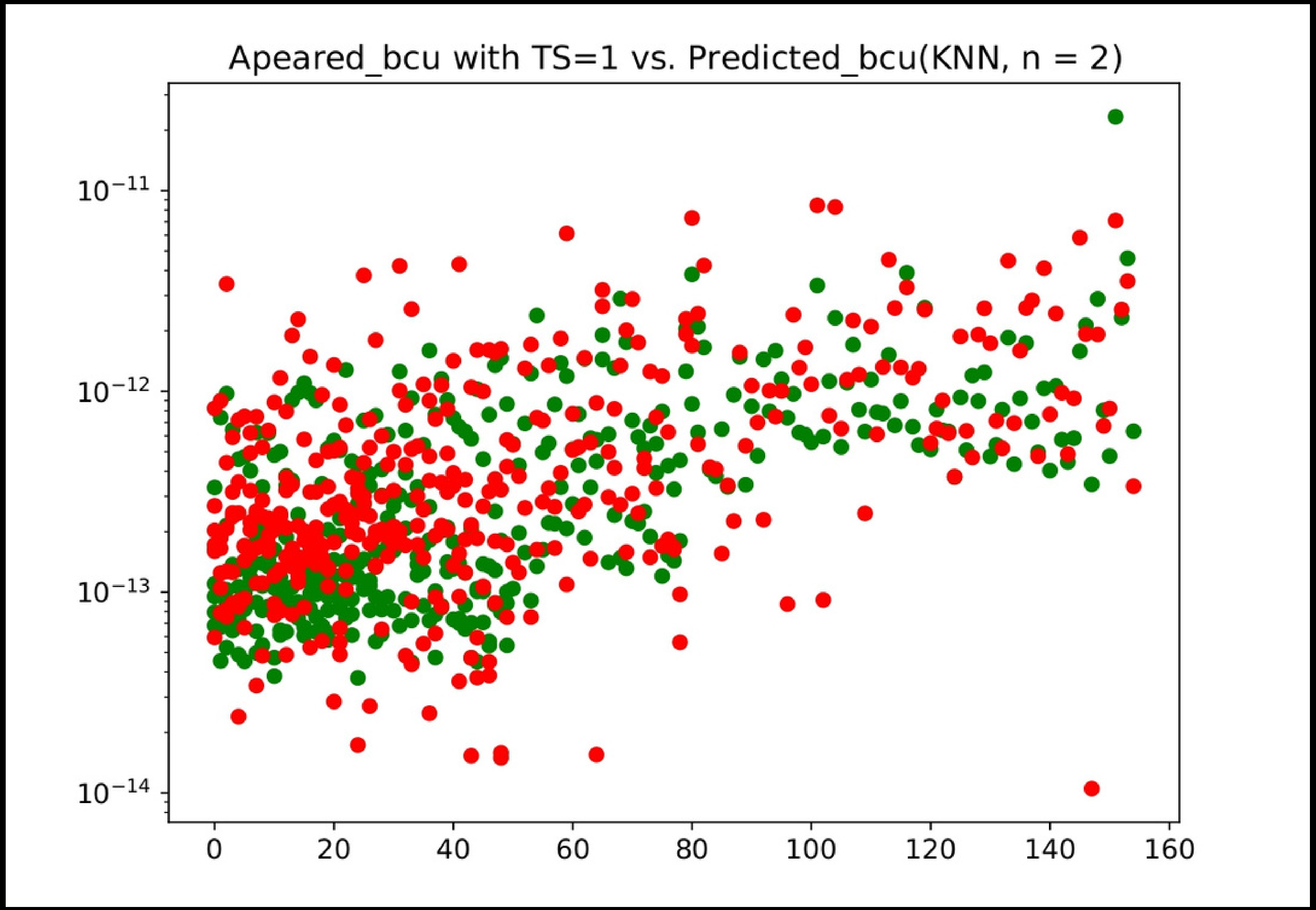
We achieved 90.06% (+/- 2.24%) validation accuracy on our data



After validation train on full dataset

Final network with 16 input neurons (10 yearly fluxes and 6 sed points) for each source, hidden layer with 256 neurons, and 2 neurons on output layer (probability of source being BLLac and FSRQ)

We achieved 92.105% Precision



Filling BCU SED data

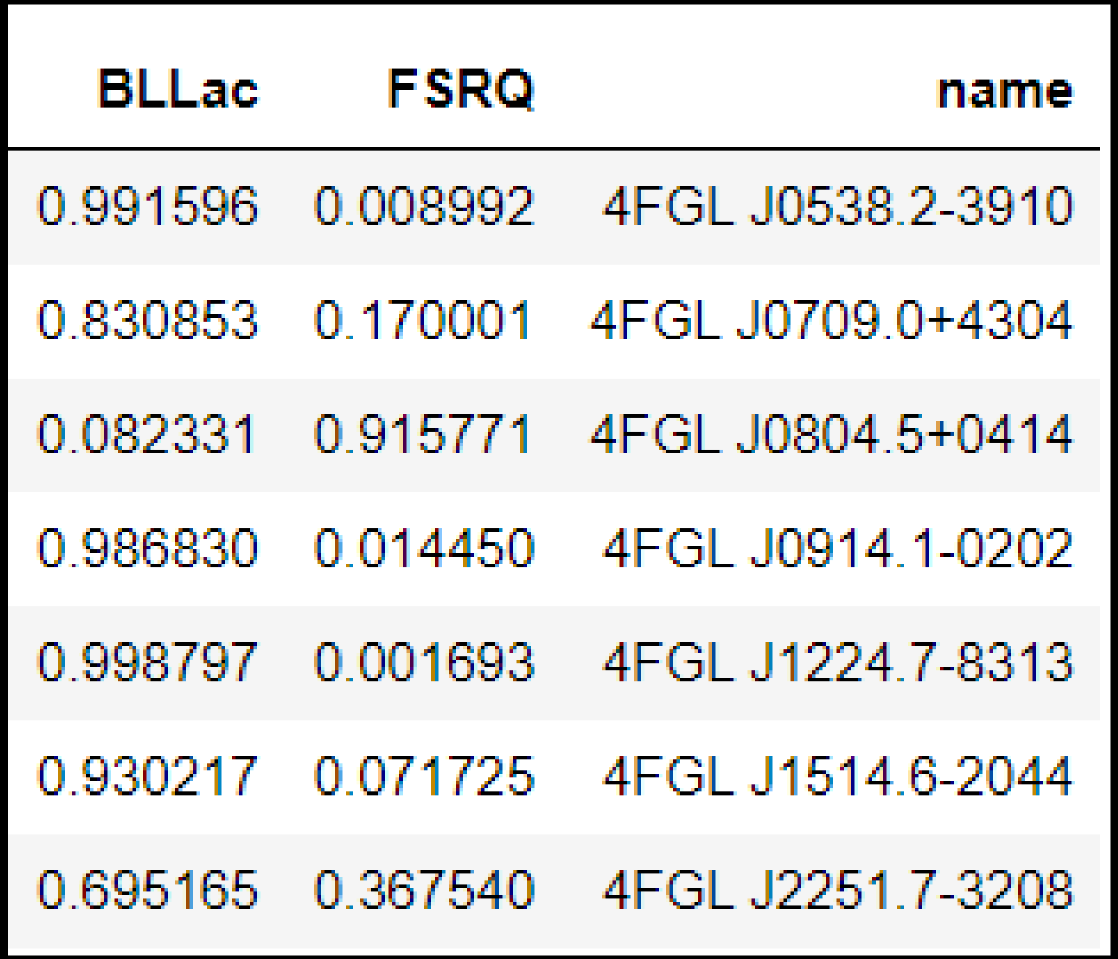
Before the classification, we must fill in the missing data of BCUs, but here we can divide only into two spectral groups, PowerLows and LogParabolas after that the imputation algorithm will find the closest pattern for a particular source and correspondingly will fill the missing data

Few predictions

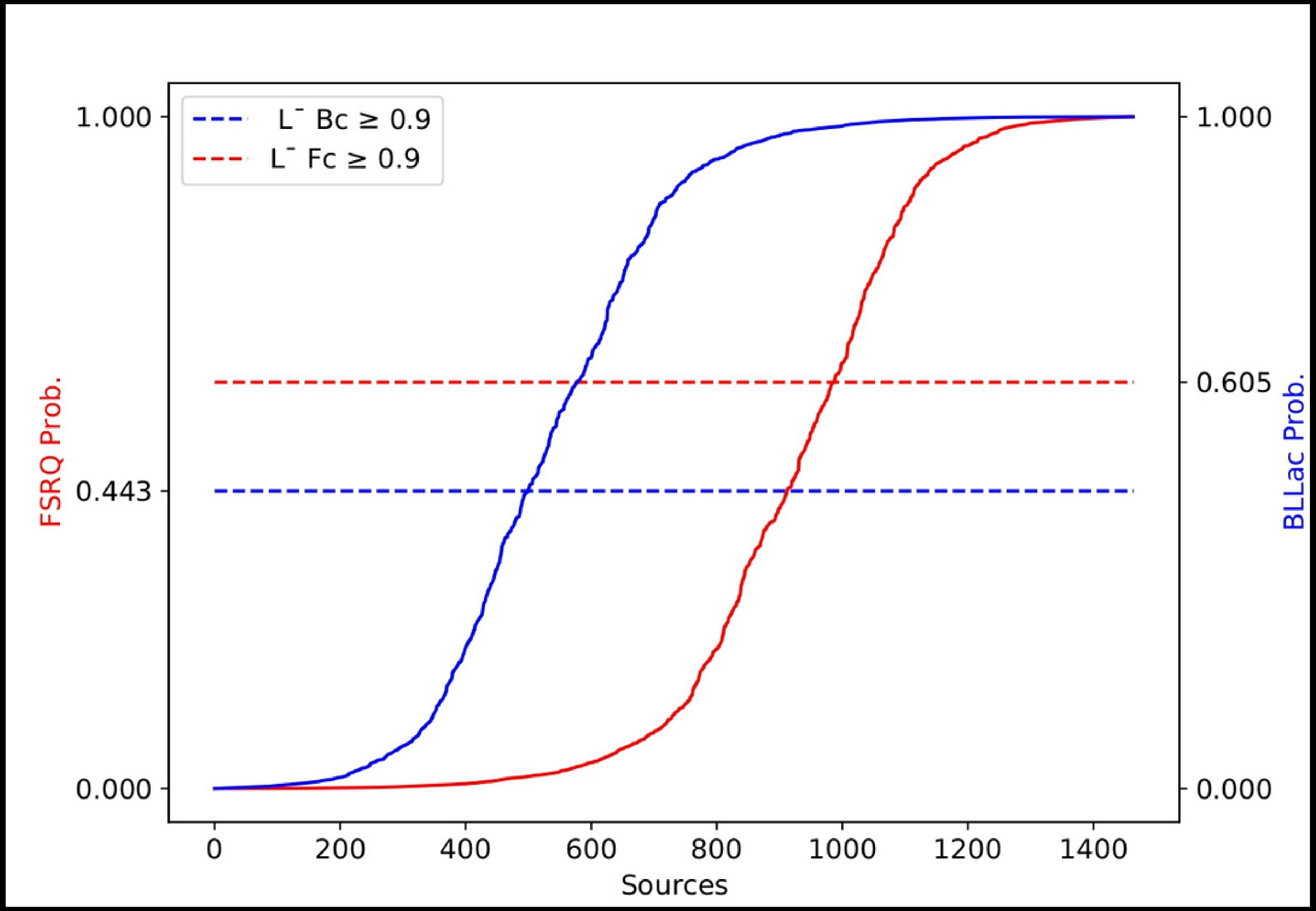
Selecting BL Lac and FSRQ candidates with a 92% precision metric

(L¯~~ Bc ≥ 0.9 and L¯ Fc ≥ 0.9; L¯ B ≥ 0.443 and L¯ F ≥ 0.605~~), 1034 BCUs are classified as BL Lacs and 426 as FSRQs, leaving 4 unclassified.

If only highly probable candidates are selected (L¯ B ≥ 0.9 and L¯ F ≥ 0.9; L¯ Bc ≥ 0.985 and L¯ Fc ≥ 0.965), then 785 BCUs are classified as BL Lacs and 278 as FSRQs. The second classification corresponds to 98 percent precision for BL Lacs and 96 percent for FSRQ.



BLLac	FSRQ	name
0.991596	0.008992	4FGL J0538.2-3910
0.830853	0.170001	4FGL J0709.0+4304
0.082331	0.915771	4FGL J0804.5+0414
0.986830	0.014450	4FGL J0914.1-0202
0.998797	0.001693	4FGL J1224.7-8313
0.930217	0.071725	4FGL J1514.6-2044
0.695165	0.367540	4FGL J2251.7-3208



BCU PROBABILITIES

The network can classify many more BCUs as almost certain BL Lacs (L¯ B → 1) than FSRQs (L¯ F → 1). This is because some BL Lacs occupy parts of parameter space where there are no FSRQs, i.e. certain group of BL Lacs are easily distinguishable from FSRQs.

— Summary

Brief summary

Our research has shown that we can develop a method that can classify blazars of uncertain types with high confidence which can be then used for studying their physical properties.

— Thank you

Your are free to ask question.