

Multi-class classification of Fermi-LAT sources with machine learning

ERLANGEN CENTRE
FOR ASTROPARTICLE
PHYSICS

Dmitry Malyshev

TeVPA 2023

Napoli, 11-15 September 2023

- Motivation for probabilistic classification of sources
- Multi-class classification of Fermi-LAT sources
- Applications for population studies

The results are mostly based on [Malyshev & Bhat \(2023\)](#)



Multiclass classification of *Fermi*-LAT sources with hierarchical class definition

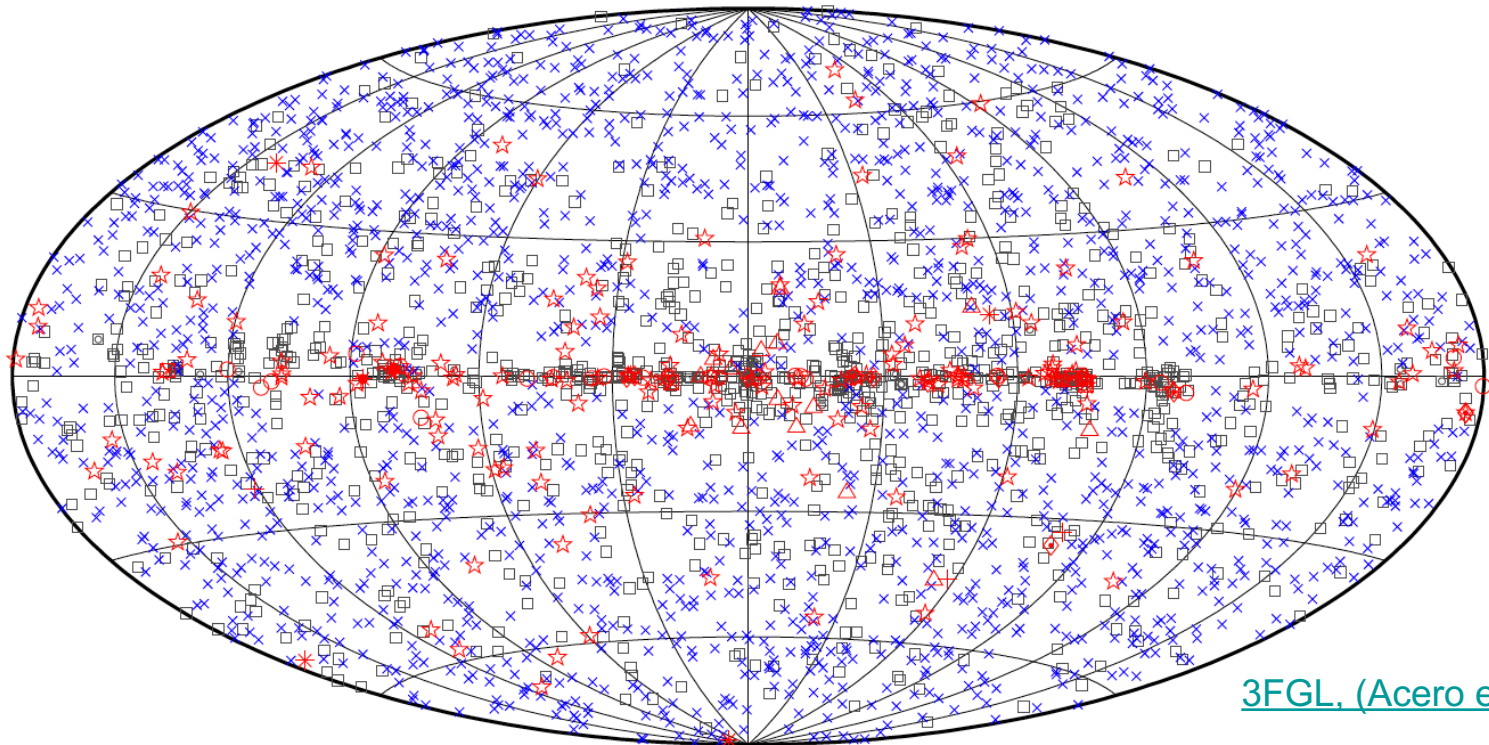
Dmitry V. Malyshev   ¹★ and Aakash Bhat ²

¹Erlangen Centre for Astroparticle Physics, Nikolaus-Fiebiger-Str 2, D-91058 Erlangen, Germany

²Institute of Physics and Astronomy, University of Potsdam, Karl-Liebknecht-Str 24/25, D-14476 Potsdam, Germany

Accepted 2023 March 26. Received 2023 March 22; in original form 2023 January 24

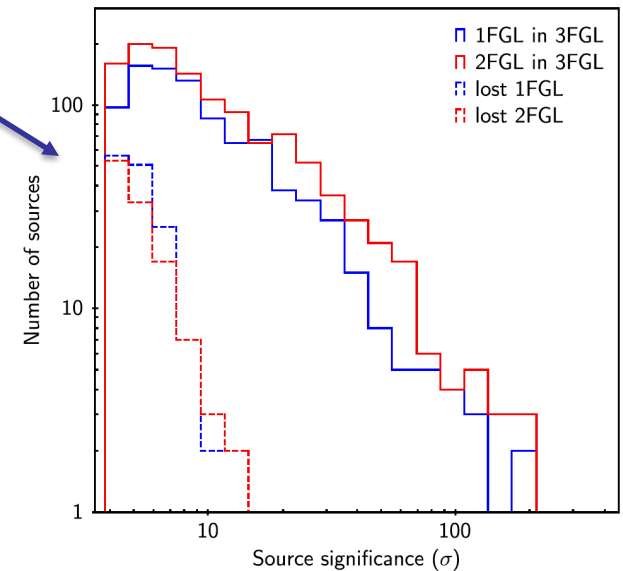
- \$200M question: what are the unassociated Fermi-LAT sources?
 - 1010 of 3033 3FGL sources ([Acero et al. 2015](#))
 - 2157 of 6658 4FGL-DR3 sources ([Abdollahi et al. 2022](#))



[3FGL, \(Acero et al. 2015\)](#)

□ No association	▣ Possible association with SNR or PWN	× AGN
☆ Pulsar	△ Globular cluster	* Starburst Galaxy
⊠ Binary	+ Galaxy	◇ PWN
★ Star-forming region	○ SNR	★ Nova

- 3FGL catalog has 3033 sources
 - 469 of the 3FGL sources are missing in 4FGL (about 15%)
 - 300/1873 (16%) of 2FGL sources are not in 3FGL
 - 310/1451 (21%) of 1FGL sources are not in 3FGL
 - These sources cannot be missing only due to statistical fluctuations
- Significance of 1FGL and 2FGL sources missing in 3FGL
- Missing sources are less significant than most of sources in the catalogs, but the significance distribution is not compatible with statistical fluctuations
- Many missing sources are “real” but incorrectly classified:
 - One source → two sources
 - Extended sources
 - Diffuse background



Acero et al. (2015)

- Classic catalog
 - Source \rightarrow association \rightarrow class (or “unassociated”)
(0, 1, 0, ..., 0)
- Probabilistic catalog
 - Source \rightarrow features \rightarrow class probabilities
(w_1, w_2, \dots, w_n)
- Advantages:
 - For unassociated sources: likely class (or classes) to which a source belongs
 - For associated sources: find outliers (sources where we are likely to make mistakes)

The problem of source classification

- In the latest Fermi LAT catalogs there are 23 classes of sources, many classes have less than 10 members

- Option 1 (mostly considered in the literature so far):

- Make two large groups of sources, e.g., Galactic and extragalactic

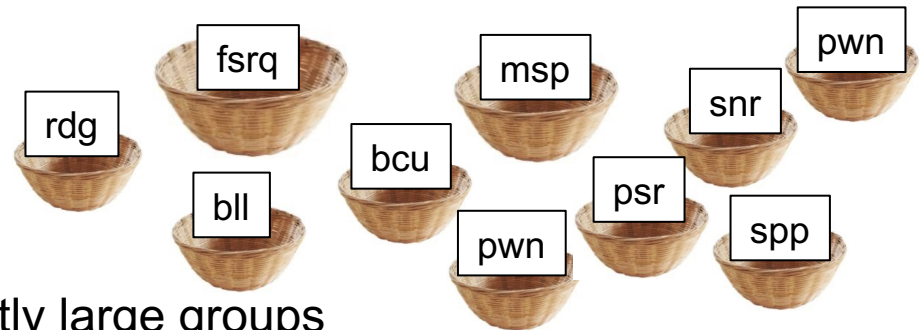
[Ackerman et al \(2012\)](#), [Saz Parkinson et al \(2016\)](#)



- Option 2:

- Take all physical classes

[Coronado-Blázquez \(2022\)](#)



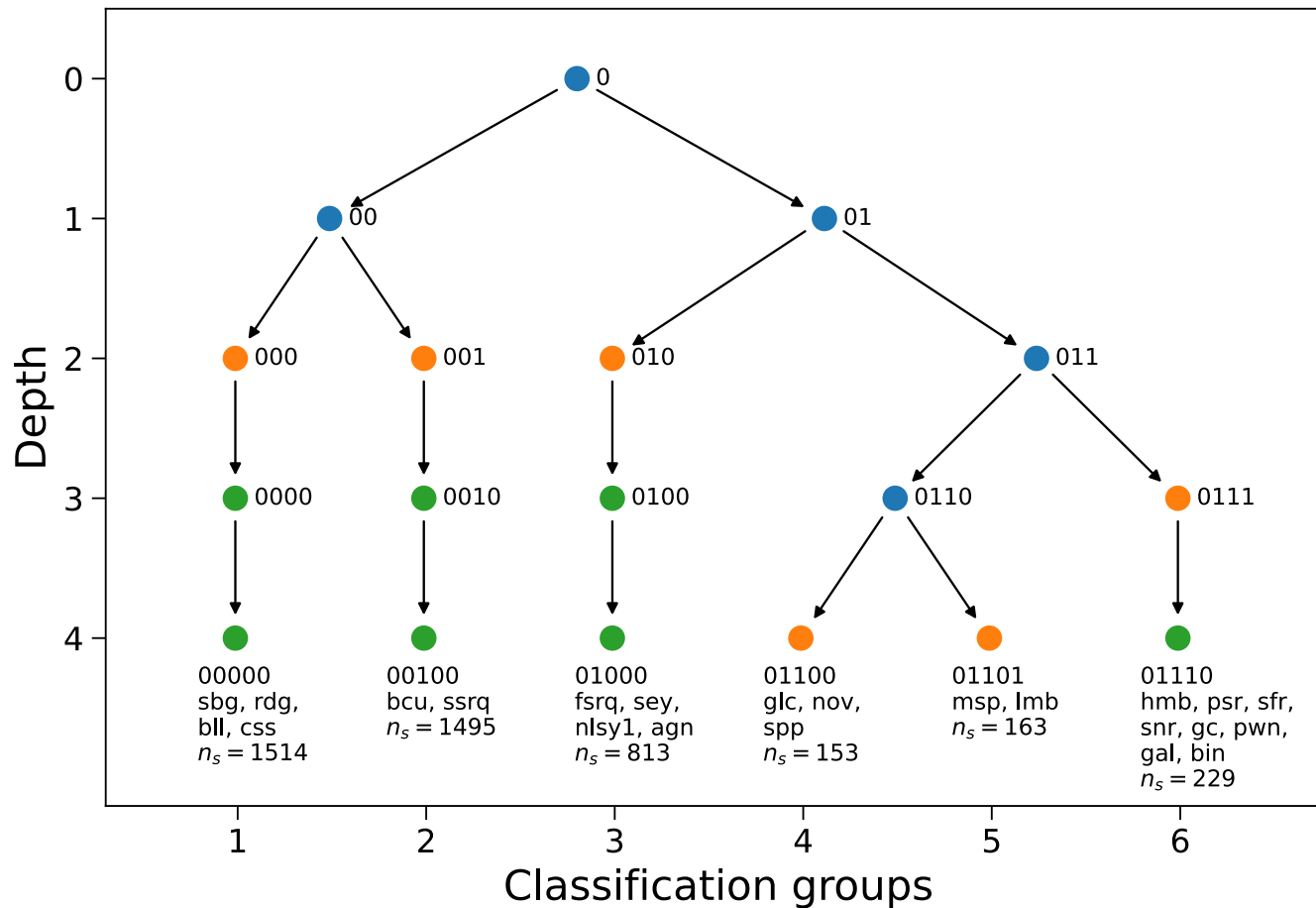
- Option 3 (“golden mean”):

- Take some number of sufficiently large groups
- Challenges:

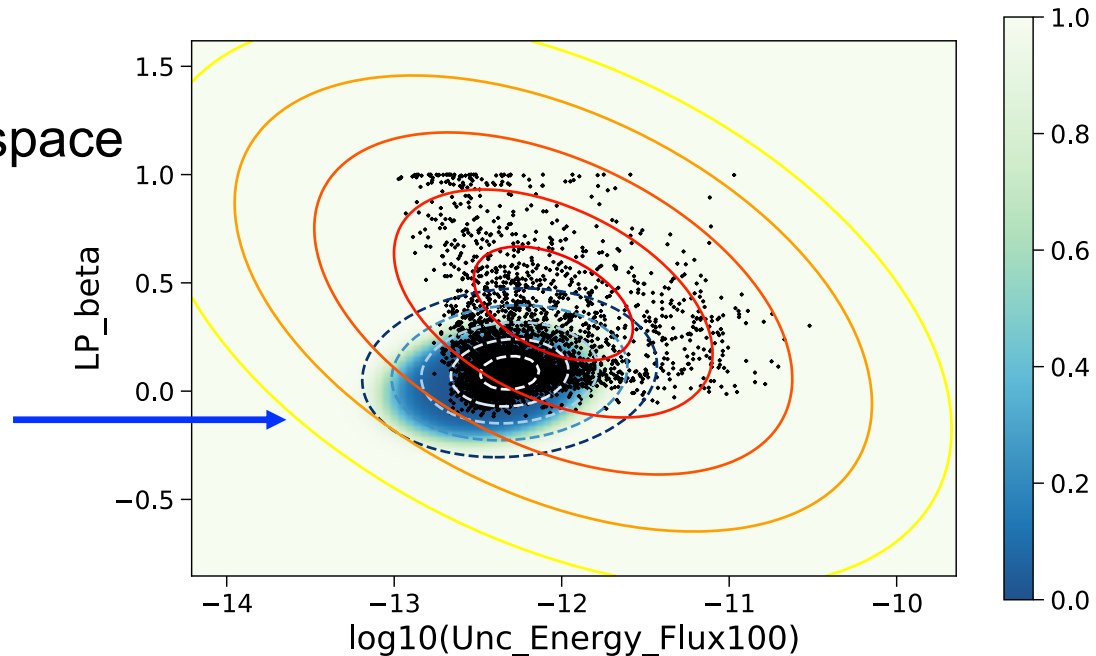
- 1) How do we define the groups?
- 2) How do we compare performance for different numbers of groups?
- 3) How do we determine a reasonable minimal number of sources in a group?

- We use the 4FGL-DR3 Fermi LAT catalog ([Abdollahi et al, 2022](#))
- 10 input features:
 - Position on the sky, energy spectrum parameters, detection significance, significance of variability in time
 - $\sin(\text{GLAT})$, $\cos(\text{GLON})$, $\sin(\text{GLON})$, $\log_{10}(\text{Energy_Flux100})$, $\log_{10}(\text{Unc_Energy_Flux100})$, $\log_{10}(\text{Signif_Avg})$, LP_index1000MeV , LP_beta , LP_SigCurv , Variability_Index
- Training and testing: identified and associated source
- Classification algorithms:
 - Random forest (RF)
 - Neural networks (NN)

- Iterative division of physical classes into groups of smaller and smaller size with each division step



- Represent the distribution of sources in the feature space as a superposition of two Gaussians
- Example of a Gaussian mixture model (GMM) for two features
- Contours show 1, 2, ..., 5 sigma levels for the two Gaussians



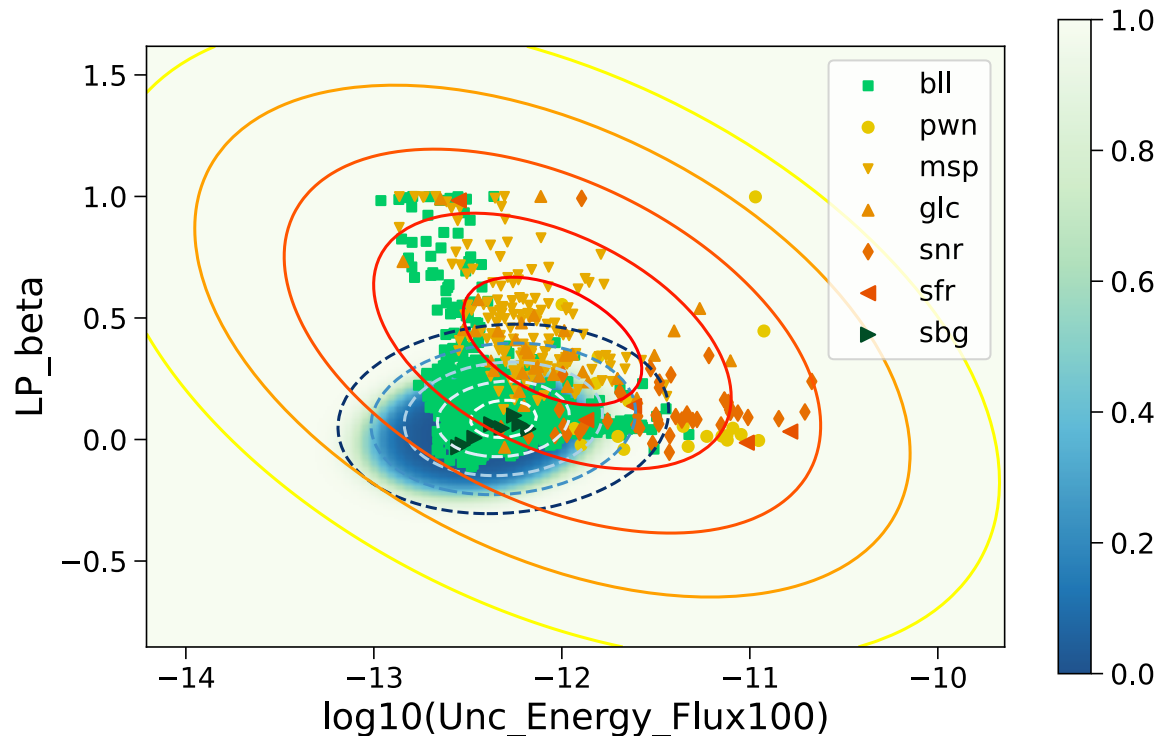
- The probability for a source with features x to belong to the distribution given by Gaussian “1” (shown by yellow color on the plot):

$$p_1 = \frac{G_1(x)}{G_0(x) + G_1(x)}$$

- The probability to belong to Gaussian “0” is $p_0 = 1 - p_1$

Class division with Gaussian mixture model

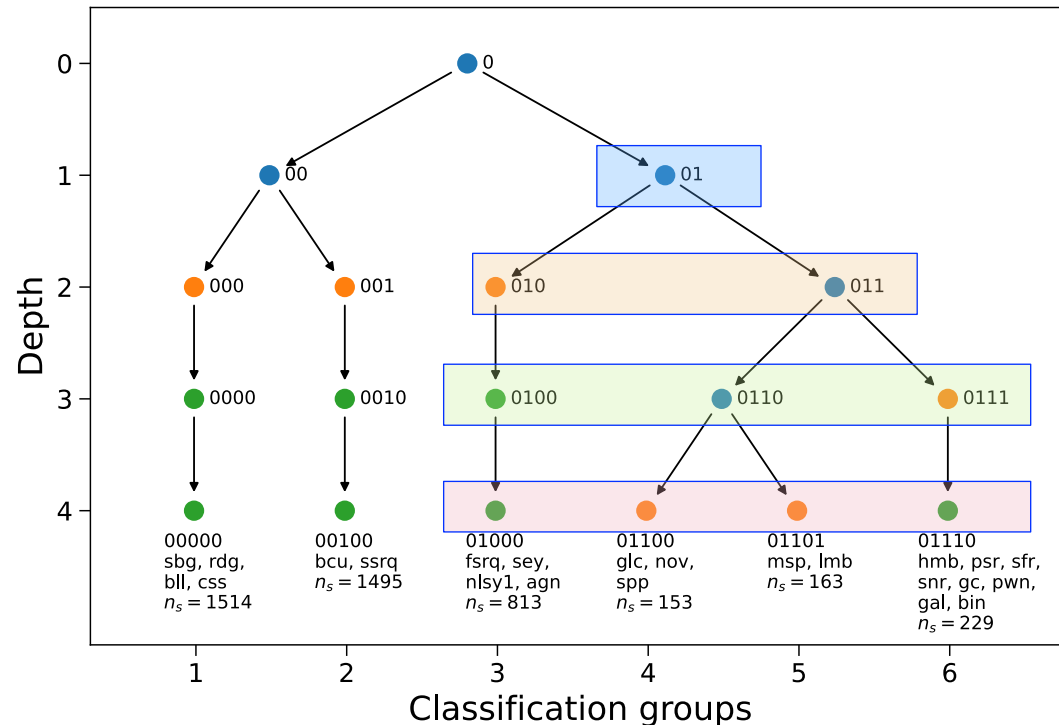
- Depending on the probabilities of sources in a physical class, we assign the class either to group 0 or to group 1.



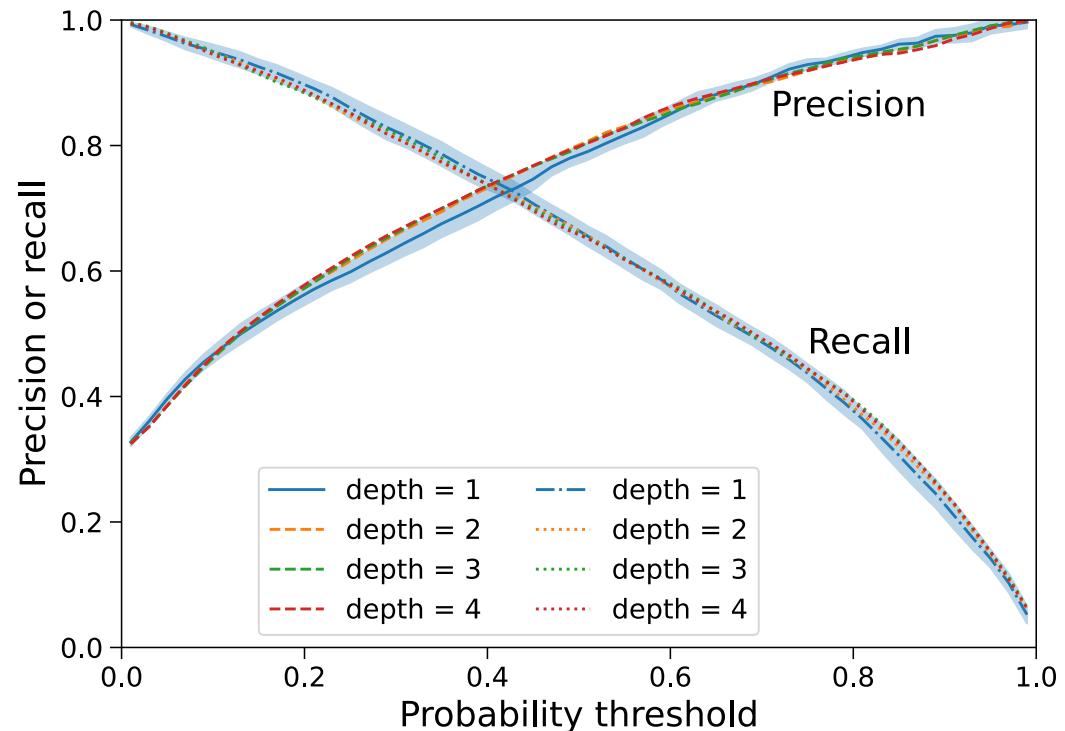
- Examples of physical classes. Group 0: **bll**, **sbg**; group 1: **pwn**, **msp**, **glc**, **snr**, **sfr**.
- Continue subdivision

Comparison of performance via tree structure

- One of the main advantages of hierarchical class definition is the ability to compare performance for different numbers of classes.
- The class probability of a source can be computed by the classification at the given depth or by adding class probabilities of children nodes.
- For example, the probability of class “01” can be estimated in four different ways by adding the class probabilities of classes in rectangles at depths 1 – 4.
- Compare the 2-class classification performance, if we compute the 2 class probabilities in 2-, 4-, 5-, or 6-class classifications.



- We calculate the precision and recall for class “01” as a function of class 01 probability calculated in 4 different ways, which correspond to the depth of the class splits.
- The precision and recall are similar for the different depths
 - Precision is slightly better for larger depths
- Band – stat. uncertainty for 2-class classification estimated from different training-testing splits (70/30%)
- We use random forest with 50 trees and max depth of 15



- One of the main questions is: how many sources in different classes, e.g., pulsars or MSPs, are among the unassociated sources?
- The expected number of sources in a class m is given by the sum of class- m probabilities over all unassociated sources.

$$N_{\text{unas}}^m = \sum_{i \in \text{unas}} w_i^m$$


- As a cross-check one can estimate the number of associated sources in class m by the sum of class- m probabilities over the associated sources and compare with the actual counts of associated sources

$$N_{\text{assoc}}^m = \sum_{i \in \text{assoc}} w_i^m$$


Numbers of sources

Number of
associated
sources

Number of associated
sources estimated with
RF or NN methods



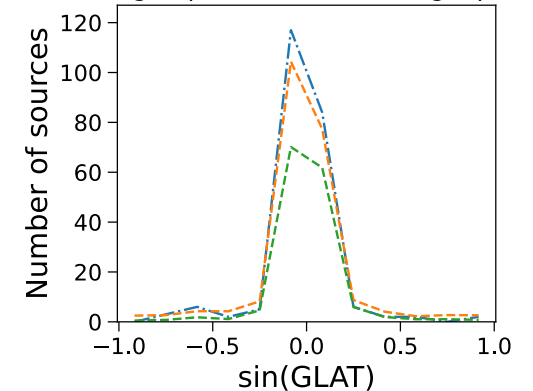
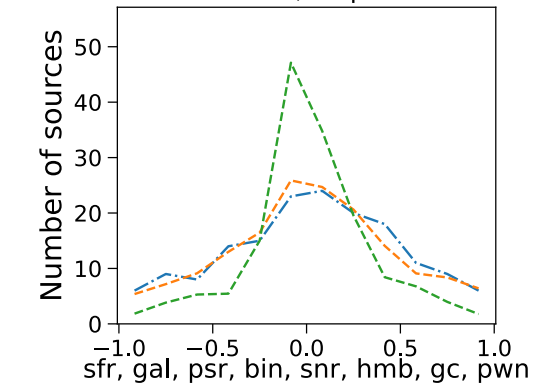
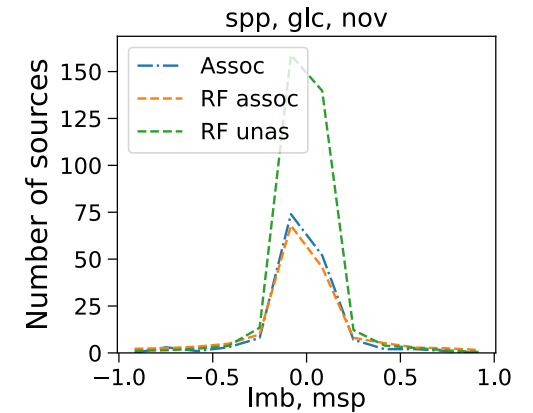
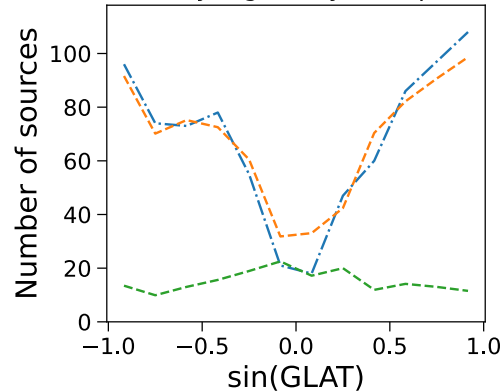
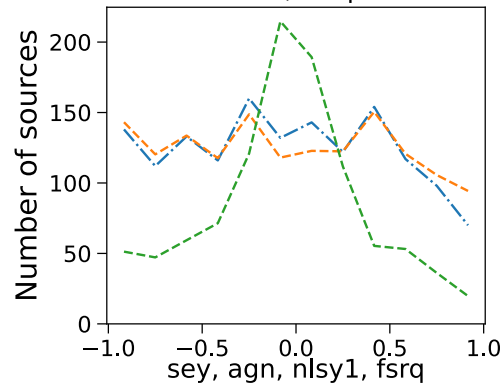
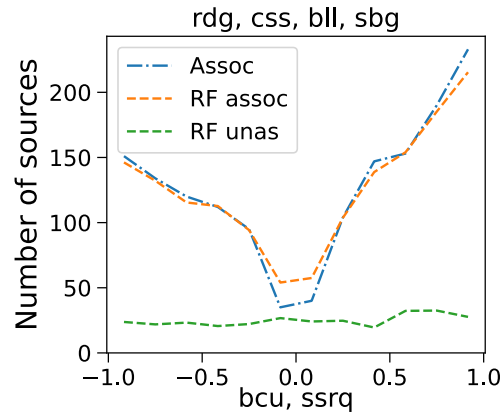
Physical classes	N assoc	RF assoc	NN assoc	RF unas	NN unas
sbg, rdg, bll, css	1514	1509.4	1529.4	312.4	267.4
bcu, ssrq	1495	1497.9	1490.8	1087.7	1099.2
fsrq, sey, nlsy1, agn	813	819.1	806.2	185.9	163.8
glc, nov, spp	153	156.4	151.9	377.4	419.8
msh, lmb	163	160.3	160.9	159.0	165.4
hmb, psr, sfr, snr, gc, pwn, gal, bin	229	224.0	227.8	168.5	175.5



Numbers of sources in
different classes among
the unassociated
sources estimated with
RF or NN algorithms

- Calculate source counts in a range of values for any variable, e.g., spectral index, flux etc.
- Here we show an example of source counts as a function of latitude bins

$$N^m = \sum_{b_1 < b_i < b_2} w_i^m$$



- Current Fermi-LAT catalogs have several limitations:
 - Many unassociated sources
 - Sources disappearing (changing identity from one catalog to another)
- Probabilistic catalogs created with ML methods provide class probabilities for both associated and unassociated sources (albeit for fewer classes than there are physical types of sources).

Use cases:

- Population studies including unassociated sources
 - Follow up studies of unassociated sources
 - Search for outliers or possible wrong associations for associated sources
- Probabilistic catalogs are available online
<https://zenodo.org/record/7538664>

Also include

- a catalog based on RF determination of classes
- a catalog with more classes: 9 classes instead of 6 classes