PROCEEDINGS OF SCIENCE

Looking for blazars among the Fermi/LAT unidentified sources in the 3FGL catalogue

Julien Lefaucheur
APC, IN2P3/CNRS – Université Paris Diderot, Paris, France
E-mail: julien.lefaucheur@apc.univ-paris7.fr

Santiago Pita
APC, IN2P3/CNRS – Université Paris Diderot, Paris, France
E-mail: pita@apc.univ-paris7.fr

Bruno Khelifi*
APC, IN2P3/CNRS – Université Paris Diderot, Paris, France
E-mail: khelifi@in2p3.fr

After 4 years of observations, the LAT telescope on board the Fermi satellite has detected the gamma-ray emission from 3034 sources (3FGL catalogue). About one third of these sources have no assigned counterparts at other wavelengths and are of unidentified nature. We present the search of blazar candidates among these unidentified sources. Considering a set of carefully choosen discriminant blazar/non-blazar parameters, we have built two multivariate classifiers (boosted decision trees and multilayer perceptron neural networks) and combined their decisions in order to obtain a high level of blazar indentification efficiency along with a well controlled level of non-blazar contamination. The low latitude and high latitude cases have been treated separately. Applying these classifiers to the sample of 1009 unidentified sources, we have determined a list of 538 blazar candidates, respectively 420 and 118 located at high ($|b| > 10^\circ$) and low ($|b| \leq 10^\circ$) latitude. The corresponding level of mis-classification is estimated to be $\sim$7 sources for the high latitude case (less than 2%) and $\sim$28 sources for the low latitude case ($\sim$24%).

The 34th International Cosmic Ray Conference,
30 July- 6 August, 2015
The Hague, The Netherlands

*Speaker.

PoS(ICRC2015)845

1. Introduction

The LAT telescope, on board the Fermi satellite, is mapping since 2008 the gamma-ray sky (above 100 MeV) with unprecedented angular resolution and sensitivity. In the recently published 3FGL catalogue [1], the Fermi/LAT collaboration reports the detection of gamma-ray emission from 3034 sources, obtained after 4 years of observations. Two third of these sources have associated counterparts, mainly active galactic nuclei (AGN) but also galactic sources such as pulsars, pulsar wind nebulæ (PWN) and supernovæ remnants (SNR). Most of the detected AGN are blazars (BL Lac or FSRQs). The understanding of their population and evolution – for example the validity of the "blazar sequence" – and the determination of the extragalactic background light (EBL) are key topics to high-energy astrophysics [2] which are currently limited, in the observational side, by the small number of detected blazars.

The aim of the work presented here was to significantly increase the number of gamma-ray blazars candidates by developing and applying efficient classification methods in order to identify new blazars among the high number of the so called Fermi/LAT unidentified sources. For that, we use the 3FGL catalogue to build different classifiers based on differences between blazars and other identified sources (pulsars, PWN, SNRs). After a review of different types of classifiers and a large set of possible discriminant parameters, we selected two classification methods and six parameters, considering the blazar/non-blazar separation performance and the stability they were able to allow.

The signal and background samples, as long as the selected discriminant parameters are described in section 2, the choice of two multivariate analyses is presented in section 3 and the determination of their operating points and the corresponding performance are presented in section 4. A summary of the results is presented in section 5.

A table containing the results of this work (blazar candidates among the unidentified sources of the 3FGL catalogue) is available on http://unidgamma.in2p3.fr.

2. Samples and discriminant parameters using the 3FGL catalogue

To build a classifier, it is necessary to define a signal sample and a background sample. For the signal sample, a total of 1717 blazars (660 BL Lac, 484 FSRQs and 573 blazar candidates) were selected among the 3034 sources of the catalogue. For the background sample, two different situations were considered, depending on the proximity of sources to the galactic plane. In the first case, for the study dedicated to low latitude sources ($|b| \leq 10°$) we used a set of 246 galactic sources containing a large number of pulsars (166), and a few tens of PWN and SNRs. In the second case, for the study dedicated to the high latitude sources ($|b| > 10°$), we followed the suggestion of [3] and considered only the population of pulsars which is in this case more likely to contaminate the blazars population. This choice tends to increase the blazar/non-blazar separation power.

In order to determine a set of discriminant parameters, we made a review of the available parameters in the 3FGL catalogue and examined also those already used in previous studies [4, 5, 3, 6]. Two types of parameters appear to be particularly interesting to discriminate between blazars and the other classes of sources: those quantifying the variability of the sources, as we expect blazars to be variable as opposed to pulsars, PWN and SNRs; and the spectral parameters, as blazar's spectra are generally well adjusted by a simple power law or a log parabola, whereas

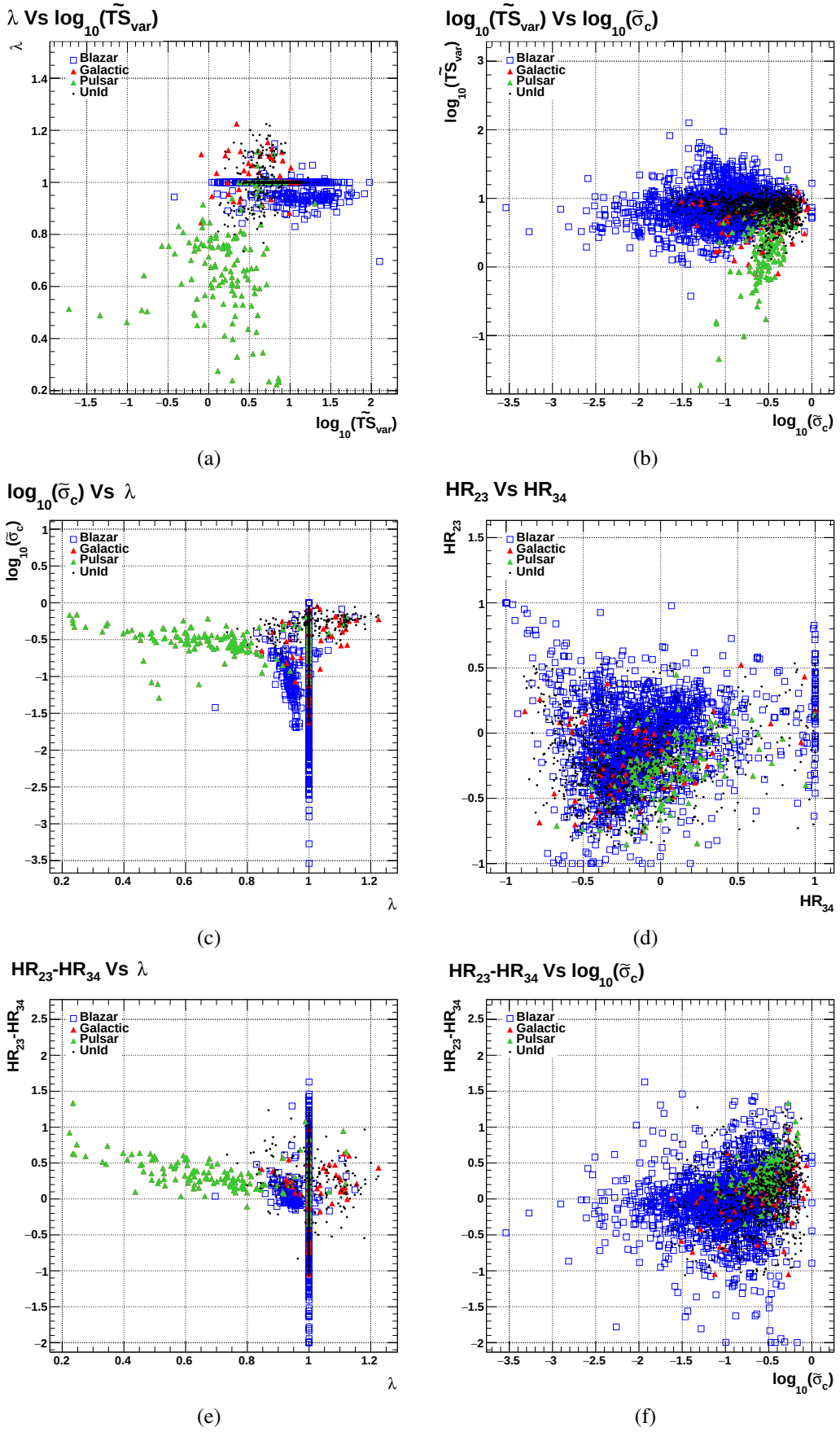

Figure 1: Scatter plots for selected discriminant variables for blazars (blue squares), sources belonging to our Galaxy (red triangles), pulsars specifically (green triangles) and the unidentified sources (black dots).

pulsars, for example, generally show a curved spectrum typically well adjusted by a broken power law or a power law with an energy cut-off. In order to minimize the biases in the final classifiers performance, we decided to discard those parameters whose separation power shows a clear dependency with the flux level of the sources. We finally selected 6 discriminant parameters, considering individually the increase of separation power and the stability that they provide to the classifiers. Five of these parameters have been already used in previous studies: $\widetilde{\sigma}_c$ defined as σ_c/σ where σ_c is the significance of the curvature and σ is the detection significance [6]; the normalised variability, called $\widetilde{\mathrm{TS}}$, given by the ratio between the index variability TS and σ [6]; the hardness ratios HR_{23} and HR_{34} as well as their difference $\mathrm{HR}_{23} - \mathrm{HR}_{34}$ [4]. Additionally, we introduced a new parameter, called λ, defined as the ratio between the spectral index of the preferred hypothesis and the spectral index of the power law hypothesis. A selection of scatter plots is shown on figure 1 for different source samples.

3. Classifications and multivariate analyses

Once the discriminant parameters are selected, a multivariate analysis needs to be chosen for the problem, here a binary classification blazar/non-blazar classification. The construction of a classification follows two steps. During the first one, a fraction of the signal and the background samples, called the training sample, is used to construct a final discriminant parameter, called ζ, maximising the signal/background differences. The second step makes use of the remaining part of the samples, called the test sample, to derive the performance, i.e. the rate of blazar identification $\varepsilon_{\mathrm{Sig}}$ and the rate of false identification $\varepsilon_{\mathrm{Bkg}}$, both as a function of ζ.

The TMVA package [7] proposes different multivariate analyses methods to tackle this kind of problem such as random forests, neural networks, support vectors machine or boosted decision trees (BDT). After having tested different settings for these methods it has quickly appeared they could reach comparable performance. The choice was made to use two of them with different nature, the boosted decision trees and a multilayer perceptron (MLP) neural network. In order to reduce significantly the false association rate, we decided to tag a source as a blazar candidate only if both classifiers agree on the blazar-like nature of a source.

For the BDT method we have chosen to generate a large forest of short trees ($ntrees = 400$, $depth = 3$). In order to reach the performance of the other classifiers, the learning algorithm differs slightly from the original AdaBoost: before the generation of a decision tree, the events of the training samples are selected n times according to a given probability following a Poisson law of parameter 0.8 ($UseBaggedBoost = true$, $BaggedSampleFraction = 0.8$). No transformations, such as decorrelation using principal component analysis or gaussianization, were applied to the input variable, since no significant improvement of the performance were observed. For the neural network method, the architecture has been set to a single hidden-layer composed of $N_{\mathrm{var}} + 15$ neurons[1]. Complexifying the architecture of the network by adding more neurons or more hidden layers didn't improve significantly the performance. The research of the minimum of the error function has been done using the backpropagation algorithm. Following the suggestion of [7], the input variables were normalised between -1 and $+1$ for the neural network. Other transformations

[1] N_{var} is the number of discriminant variables used to build up the classification.

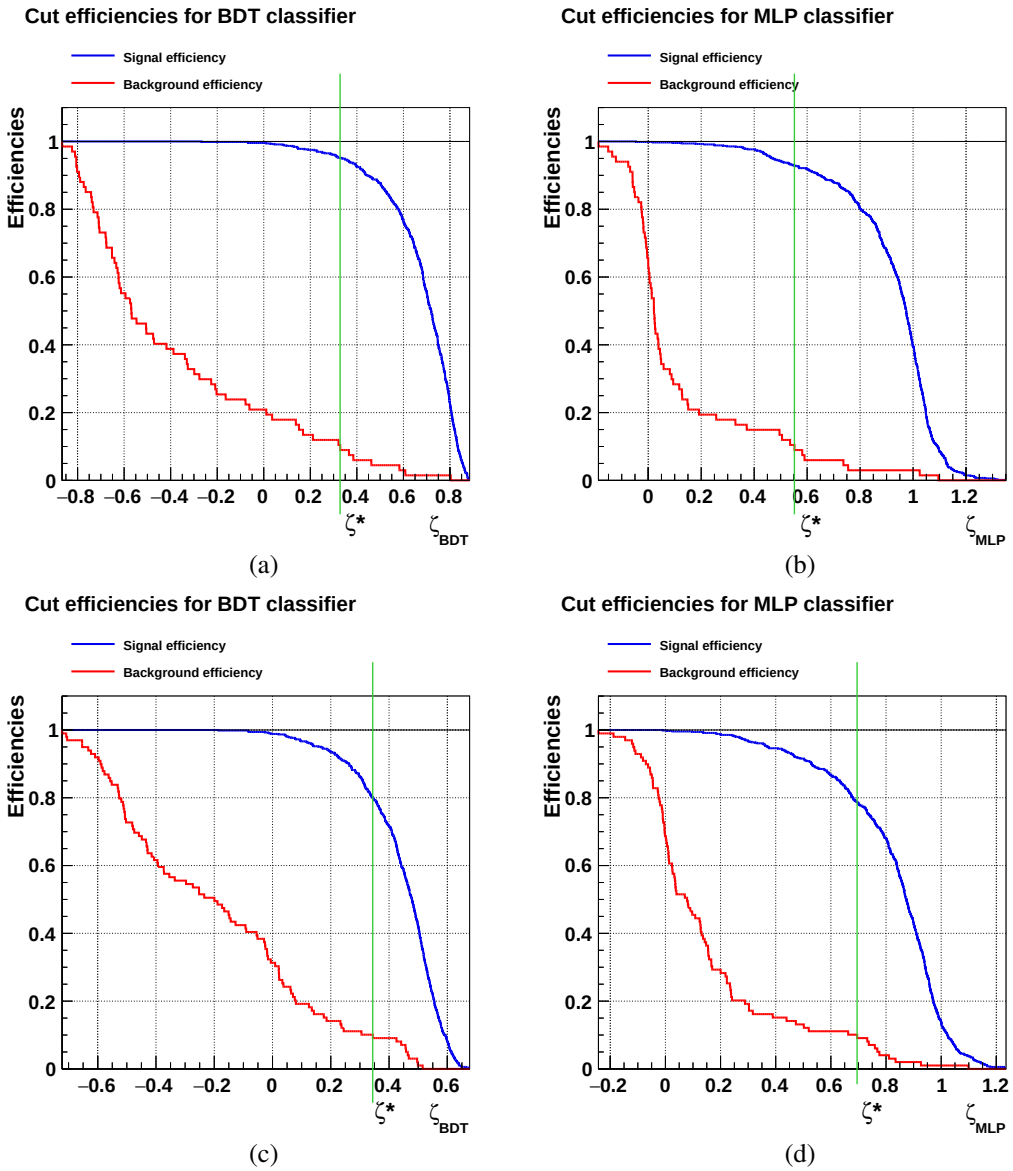

Figure 2: Signal (blue) and background (red) efficiencies as a function of ζ for the BDT (left) and MLP (right) classifiers, corresponding to the high galactic latitude (upper panel) and the low galactic latitude (lower panel). In each case, the value of $\zeta^{\star}$ is indicated by the green line.

didn't add more background rejection. Finally, as the signal and background samples have different sizes, we normalized the events in order to have samples with identical sizes[2] ($NormMode = EqualNumEvents$).

[2]For BDT, this is naturally done with the AdaBoost algorithm.

		N_s	N_c	N_{lost}	N_{bad}
High latitude region	GF	422	336	29	5
	BF	119	84	7	2
Low latitude region	GF	169	50	14	9
	BF	299	68	17	19

Table 1: Summary of the results obtained when applying the classifiers to the high and low latitude unidentified sources according to their flags (GF no flag, BF flagged sources). N_s and N_c are the numbers of unidentified sources and blazar candidates, respectively. N_{lost} is the expected number of blazars lost by our classifiers and N_{bad} is the expected number of galactic sources contaminating our sample of blazar candidates.

4. Operating point and performance

To evaluate the average performance of both classifiers we performed 300 training and test phases for both the BDT and the MLP methods. For each of these, 60 % of the training sample was randomly chosen for the learning phase and the remaining part was used to derive the performance. To define the operating point for each method, a simple approach was considered: we imposed $\varepsilon_{Bkg} = 10\,\%$ for both classifiers and we found the corresponding values of $\zeta^\star$ and then ε_{Sig}, the first one defining the threshold on the discriminant variable above which one source is considered as a blazar candidate by each classifier.

For the analysis dedicated to the study of the high latitude sources, the BDT and the MLP methods allow to get a blazar identification rate of $(93.9 \pm 2.5)\,\%$ and $(94.1 \pm 2.0)\,\%$, respectively (errors are statistical). In order to strengthen the decision power, we combine both decisions (keeping only sources which are considered as blazar candidates by both classifiers). The combined performance is then given by $\varepsilon_{Sig} = (91.8 \pm 2.6)\,\%$ and $\varepsilon_{Bkg} = (6.8 \pm 1.2)\,\%$. Similarly, for the analysis dedicated to the study of low latitude sources, the BDT and the MLP methods allow to get a blazar identification rate of $(82.2 \pm 5.3)\,\%$ and $(82.0 \pm 5.7)\,\%$, respectively. The combined decision allows to reach $\varepsilon_{Sig} = (77.3 \pm 6.2)\,\%$ and $\varepsilon_{Bkg} = (7.7 \pm 1.0)\,\%$.

Finally, for each region of the sky, two single classifiers (BDT and MLP) were selected to be further used to identify blazar candidates among the 3FGL unidentified sources. We select them on the basis of the compatibility of their performance (individual and combined, see figure 2) with the average behaviour described above. We also verified the absence of overtraining effects by comparing the ζ distributions of the training and test samples (Kolmogorov-Smirnov test).

5. Results

Once the technical choices are fixed and a classifier is selected for each BDT and MLP method, and for each high ($|b| > 10^\circ$) and low ($|b| \leq 10^\circ$) latitude analyses, we apply them once to the sample of 1009 unidentified sources according to their positions in the sky. This results in a total of 538 blazar candidates. Results are summarized in table 1 and the entire list of blazar candidates is available on http://unidgamma.in2p3.fr. A short sample is shown on table 2. We found 420 blazar candidates at high latitude and 118 close to the galactic plane. Among them, respectively 20 %

3FGL name	l (°)	b (°)	$\widetilde{\sigma}_c$	$\widetilde{\mathrm{TS}}$	HR_{23}	HR_{34}	$\mathrm{HR}_{23}-\mathrm{HR}_{34}$	λ	ζ_{BDT}	ζ_{MLP}
J0000.2-3738	345.411	-74.947	0.29	8.81	0.55	0.23	0.32	1.00	0.6954	1.0227
J0002.0-6722	310.139	-49.062	0.23	7.83	-0.00	0.12	-0.12	1.00	0.7084	0.9272
J0004.2+0843	103.599	-52.363	0.17	9.61	0.06	0.52	-0.45	1.00	0.8023	1.0911
J0006.2+0135	100.401	-59.297	0.03	8.25	-0.05	-0.20	0.14	1.00	0.7543	0.9769
J0006.6+4618	114.908	-15.867	0.20	10.60	-0.37	-0.32	-0.04	1.00	0.7398	0.9615
J0007.4+1742	108.332	-43.911	0.17	10.21	-0.15	-0.24	0.09	1.00	0.7463	0.9339
J0007.9+4006	113.977	-22.007	0.22	9.82	-0.44	0.07	-0.51	1.00	0.7770	0.9933
J0010.5-1425	84.006	-74.112	0.66	13.16	-0.25	-0.42	0.17	1.00	0.5472	0.7557
J0016.5+1713	111.141	-44.850	0.55	12.94	-0.19	-0.76	0.57	1.00	0.5051	0.8998
J0017.1+1445	110.672	-47.293	0.25	7.54	0.14	-0.43	0.57	1.00	0.4979	0.5884
...	...	...	...	...	...	...	...	...	...	...

Table 2: First ten blazar candidates with no "analysis flag" in the high latitude region ($|b| > 10°$). The columns correspond respectively to the 3FGL name, the galactic coordinates of the source (l,b), the values of the 6 discriminant parameters, and finally the values of ζ builded with the BDT and MLP classifiers. This table is available in its entirety on http://unidgamma.in2p3.fr.

and 58 % have an "analysis flag" in the Fermi/LAT catalogue, indicating that a possible problem has arisen during the reconstruction of the γ-ray sources. Knowing $\varepsilon_{\mathrm{Sig}}$, $\varepsilon_{\mathrm{Bkg}}$ and the number of blazar candidates, we estimate that our high and low latitude samples of blazar candidates are contaminated by $\sim$7 and $\sim$28 galactic sources, respectively, and that we missed $\sim$36 and $\sim$31 blazars, respectively. These numbers should be considered as estimations, in particular because the systematics related to the presence of sources with a Fermi/LAT "analysis flag" both in the training sample and in the sample of unidentified sources have not been fully estimated.

The work presented here is based on the coupling of the decisions of two classifiers built on a set of parameters carefully selected on the basis of their blazar/non-blazar separation power. It is, as far as we know, the first study of this kind based on the 3FGL catalogue. A possible extension could be the determination of the nature of the blazar candidates using the BL Lac/FSRQs differences imprinted in the 3FGL catalogue. Beyond that, a multiwavelength approach is necessary to firmly confirm the nature of the blazar candidates proposed in this work.

Acknowledgement

We would like to thank Catherine Boisson (LUTh, Observatoire de Paris) and Arache Djannati-Ataï (APC, IN2P3/CNRS) for useful discussions at different levels of this work. We also acknowledge the financial support of the APC laboratory.

References

[1] The Fermi-LAT Collaboration, *Fermi Large Area Telescope Third Source Catalog*, *ArXiv e-prints* (Jan., 2015) [arXiv:1501.0200].

[2] H. Sol et al., *Active Galactic Nuclei under the scrutiny of CTA*, *Astroparticle Physics* **43** (Mar., 2013) 215–240, [arXiv:1304.3024].

[3] N. Mirabal et al., *Fermi's sibyl: mining the gamma-ray sky for dark matter subhaloes*, MNRAS **424** (jul, 2012) L64–L68.

[4] M. Ackermann et al., *A Statistical Approach to Recognizing Source Classes for Unassociated Sources in the First Fermi-LAT Catalog*, ApJ **753** (July, 2012) 83, [arXiv:1108.1202].

[5] E. C. Ferrara et al., *Fermi's Mystery Sources: Methods for Classification and Association*, *ArXiv e-prints* (June, 2012) [arXiv:1206.2571].

[6] M. Doert et al., *Search for Gamma-ray-emitting Active Galactic Nuclei in the Fermi-LAT Unassociated Sample Using Machine Learning*, ApJ **782** (Feb., 2014) 41, [arXiv:1312.5726].

[7] A. Hoecker et al., *TMVA - Toolkit for Multivariate Data Analysis*, *ArXiv Physics e-prints* (Mar., 2007) [physics/0703039].