PROCEEDINGS OF SCIENCE

ONLINE ICRC 2021
THE ASTROPARTICLE PHYSICS CONFERENCE
Berlin | Germany
37th International Cosmic Ray Conference
12–23 July 2021

Searching for dark matter in Fermi-LAT unidentified sources with Neural Network

Viviana Gammaldi,[a,*] **J. Coronado-Blázquez,**[a] **M.A. Sánchez-Conde**[a] **and Bryan Zaldivar**[a,b]

[a] *Departamento de Física Teórica & Instituto de Física Teórica UAM/CSIC, Universidad Autónoma de Madrid, E-28049 Madrid, Spain*

[b] *Instituto de Fısica Corpuscular, Universidad de Valencia and CSIC*

E-mail: viviana.gammaldi@uam.es

Around one third of the point-like sources in the *Fermi*-LAT catalogs remain as unidentified sources (UniDs) today. Indeed, these unIDs lack a clear, univocal association with a known astrophysical source identified at other wavelengths, or to a well-known source type emitting only in gamma rays (such as certain pulsars). If the dark matter (DM) is composed of weakly interacting massive particles (WIMPs), there is the exciting possibility that some of these unIDs may actually be DM sources, emitting gamma rays by WIMPs annihilation. We propose a new search methodology that uses Machine Learning classification algorithms calibrated to a mixed sample of both experimental (known astrophysical objects) and theoretical (expected DM) data. With our methodology, we can correctly classify a promisingly high percent of astrophysical sources, opening a window to robustly search for DM source association among *Fermi*-LAT unIDs.

37th International Cosmic Ray Conference (ICRC 2021)
July 12th – 23rd, 2021
Online – Berlin, Germany

*Presenter

PoS(ICRC2021)493

1. Introduction

In recent years, more and more Machine Learning (ML) algorithms have be applied to a broad variety of open questions in both physics and astrophysics [1]. Among other applications in the astrophysical and cosmological context, the *Fermi*-LAT satellite [2] provides a nice laboratory for ML application, due to the open-access policy of data sharing, that are provided on-line to scientific community. Indeed, very interesting, around one third of the point-like gamma-ray sources in the 4FGL *Fermi*-LAT catalogs [3] remain as unidentified (unIDs) today. These unIDs lack a clear, univocal association with a known astrophysical source identified at other wavelengths, or to a well-known source-type emitting only in gamma rays (such as certain pulsars). Many efforts have been already devoted in to apply classification algorithms to such a catalogue of gamma-ray data, e.g. by constructing probabilistic catalogues of *Fermi*-LAT unIDs (see e.g. [4] and references therein). Most of these works focus on classifing unIDs as different types of known astrophysical sources (e.g Active Galactic Nuclei, pulsars, blazars) [5–8]. Nonetheless, if dark matter (DM) is composed of Weakly Interacting Massive Particles (WIMPs), there is also the exciting possibility that some of these unIDs may actually be DM sources, emitting gamma rays by WIMPs annihilation [9]. In fact, the nature of DM still represents an open question in physics and cosmology, and many efforts have been devoted to understand the nature of such an unknown constituent of the Universe. Those efforts include the application of different ML techniques in several related fields (e.g. [10]). In this work, we propose a novel search methodology that uses classification algorithms calibrated to a mixed sample of both experimental (known astrophysical objects) and theoretical (simulated DM) data in a derived parameter space, namely what we call the "DM-β" plot, which is defined by both the pivot energy ($E_{\rm peak}$) and the curvature (β) of the gamma-ray spectra. With our methodology, we can correctly classify a promisingly high percent of astrophysical sources, opening a window to robustly search for DM source association among *Fermi*-LAT unIDs.

This proceedings is organized as follow: in Section 2.1 we introduce the methodology adopted in this work (and inspired by [11]), in order to create a theoretically-based DM data set, which is introduced in the experimental parameter space - i.e. the so-called *Fermi*-LAT β-plot [11]. Furthermore, we introduce two *synthetic* features for DM, that are the detection significance σ_{TS} and the relative uncertainty on the curvature β_{rel}. In Section 3 we train different ML algorithms both on the benchmark two-features "DM-β" plot and the new four-features parameter space (which includes the *synthetic* ones). We find out that we can improve the precision of different ML classification algorithms (here, Logistic Regression and Artificial Neural Network) by including these *synthetic* features. In section 4 we show preliminary results of the classification of the *Fermi*-LAT unIDs as prospective DM sources, by applying our best algorithm. The conclusions are traced in Sec. 5.

2. Methodology

2.1 The "DM-β plot"

The Large Area Telescope (LAT), onboard the *Fermi* satellite [2], has revolutionized the field of gamma-ray astrophysics since its launch in 2008. Still in operation, *Fermi*-LAT is a pair conversion telescope capable to observe gamma-ray photons from energies ~20 MeV to more than 300 GeV.

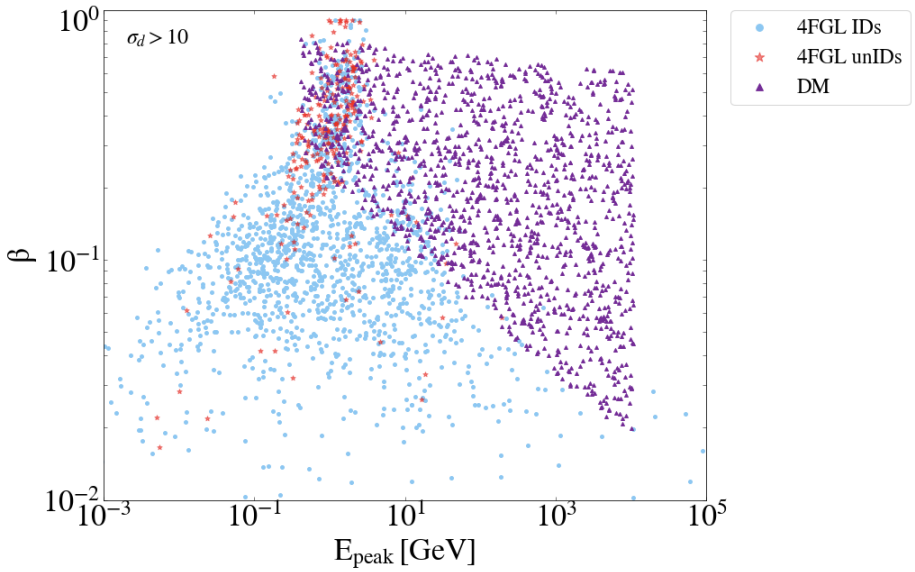

Figure 1: The "DM-β plot", which includes information about the gamma-ray spectra of well-know astrophysical gamma-ray sources (blue-light data), detected but unIDs sources (red data) and theoretical DM data set (magenta points).

Several point-source catalogs have been released and contain thousands of gamma-ray objects, many of them previously unknown. The recent 4FGL *Fermi*-LAT catalogue [3] is a collection of sources with an associated gamma-ray spectra. The latter can be generally fitted by a Log-Parabola (LP) [3]

$$\frac{dN}{dE} = N_0 \left(\frac{E}{E_0}\right)^{-\alpha-\beta\cdot\log(E/E_0)}, \tag{1}$$

which spectral features, i.e the curvature β and the pivot energy $E_{peak} = E_0 \cdot e^{\frac{2-\alpha}{2\beta}}$, represent a signature of different kind of emitting sources. Indeed, different astrophysical sources - as well as detected unIDs - occupy different regions in the so-called β-plot, shown in Fig.1. In this plot, the light-blue points are astrophysical gamma-ray sources, while the red points are detected unIDs.

Similarly, we can predict the gamma-ray spectra of a DM annihilation event by means of Monte Carlo event generator softwares (e.g. [12]). In fact, WIMPs annihilate in different standard model (SM) channels, which hadronization and decay processes generate spectra that are footprints of both the annihilation channel and the energy of the event, i.e. a signature of the DM candidate. The simulated spectra can be also fitted - as a first approximation - by a LP. Let us remark that the magenta DM-cloud in Fig. 1 has been created not only with the $(E_{\rm peak}, \beta)_{\rm DM}$ values obtained by a model independent hypothesis (i.e. WIMPs annihilating in a single SM channel), but also considering that DM particles may annihilate in two different SM channels, being the resulting gamma-ray spectra dN/dE given by

$$\frac{dN}{dE} = B_r \left(\frac{dN}{dE}\right)_{SM_1} + (1 - B_r) \left(\frac{dN}{dE}\right)_{SM_2} \tag{2}$$

where the branching ratio B_r is the probability of a WIMP to annihilate into one of the two SM channels. This originates different signatures in the spectra. Thus, we consider different

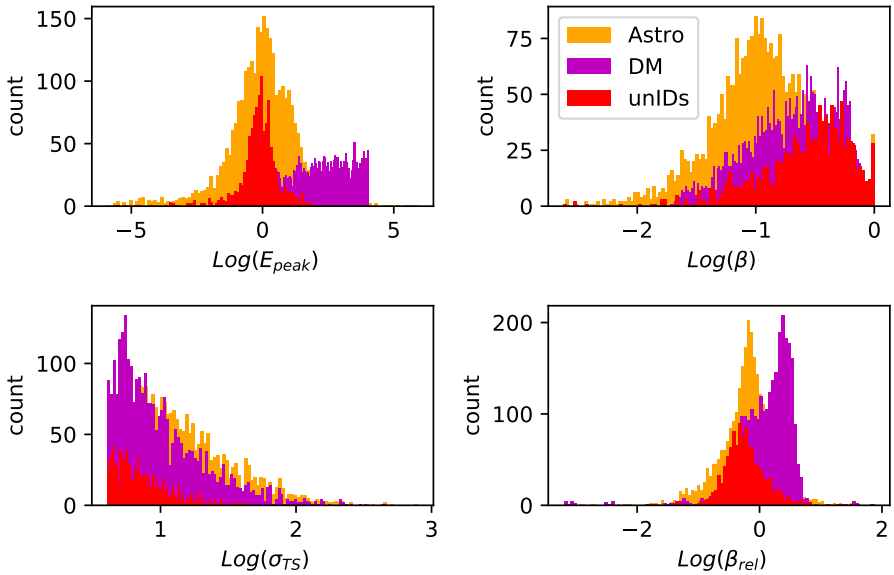

Figure 2: Histograms of the four features of the balanced data adopted in this analysis (NN), namely characteristic emission energy E_{peak} (upper left panel), curvature of the spectra β (upper right panel), detection significance σ_{TS} (lower left panel), relative error on β (lower right panel). In each panel we show the histograms for the classified astrophysical sources (yellow), unIDs (red) and DM data set (magenta).

combination of two channels with different B_r, identifying a well defined region of the "DM-β" plot (magenta points in Fig. 1) by means of the $(E_{\rm peak}, \beta)_{\rm DM}$ values obtained by the fit of the resulting spectra from [13] with the LP in Eq. (1).

2.2 *Synthetic* features

In the previous section we have introduced a novel methodology in order to introduce the WIMPs candidates in the β-plot parameter space, which allows us to distinguish and classify prospective DM candidates from astrophysical sources, based on their gamma-ray spectra. Nonetheless, this kind of training with the only two features of the "DM-β" plot represents a limitation in the framework of ML in general and Neural Network (NN) in particular. In fact, the collection of detected sources we aim to classify (here, unIDs) includes a number of features that are not considered in the theoretical data set. Among other experimental features that we are not able to "invent" (e.g. position in the sky, time variability, etc.), at least we can model part of the systematic uncertainty that are naturally included in the experimental data of detected sources (both associated sources and unIDs), which are not (yet) considered in the theoretical DM data set. In other words, we have only two features for the theoretical sample ($E_{\rm peak}$ and β), although the unIDs data include much more information themselves. In order to reduce this limitation, which reduces the precision of our classification goal (see Sec. 3), we create and include two *synthetic* features for the DM data set, i.e. the detection significance ($\sigma_{\rm TS}$) and the relative error on the spectral index ($\beta_{\rm rel} = \epsilon_\beta/\beta$). Generally speaking, *synthetic* features are additional features constructed by existing real features in order to improve the prediction of the model (see e.g. [14]). Within the hypothesis that prospective DM sources could be part of the unIDs catalogue (but they should not be included in the catalogue

of well-known astrophysical sources), these synthetic features are created in order to reproduce the distribution of the statistical significance ($\sigma_{\rm TS}$) and the uncertainty on β of detected unIDs ($\beta_{\rm rel}$). In fact, these two features are related with the sensitivity of the instrument to a certain energy, i.e. to the sistematic uncertainty associated with the detected spectra. In Fig. 2 we show the histograms of each feature for the astrophysical data set (yellow histogram), the DM data set (magenta histogram) and the unIDs (red histogram).

3. Classification accuracy

We train different classification tools available in the Scikit-Learn [15] with both the two features (2F), which define the "DM-β" plot (i.e. $E_{\rm peak}$ and β) and four features (4F) that includes the *synthetic* features (i.e. σ_{TS} and β_{rel}). In order to estimate the precision of each classifier with different features we calculate: the Overall Accuracy (OA), i.e. the number of correctly classified data set (normalized to the total number of samples in the data set): the True Negative (TN), i.e. the number of correctly classified astrophysical sources, and True Positive (TP), i.e. the number of correctly classified DM sources (both normalized to the true values, i.e. the row of each class) (see [15] for details). Preliminary results for both the Logistic Regression (LR) and NN classifiers (see Appendix A for technical details), are presented in Tab. 1. Intuitively, we trust more in the correct classification of already well-know astrophysical sources (TN) than in the correct classification of prospective DM sources (TP), i.e. our best classifier will be the one that maximizes not only the OA, but also the TN percentage (with respect to TP). Indeed, we find out that our best classifier so far is the NN applied to four features. We got $OA = 93.1\% \pm 0.4\%$ and we can correctly classify $94.7\% \pm 1.1\%$ of astrophysical sources[1].

Summary table			
LR	OA(%)	TN (%)	TP (%)
2F	84.9 ± 0.6	85.4 ± 1.3	84.4 ± 1.0
4F	86.0 ± 0.5	86.8 ± 1.2	85.6 ± 0.7
NN			
2F	86.8 ± 0.3	86.4 ± 2.4	87.2 ± 2.3
4F	93.1 ± 0.4	94.7 ± 1.1	91.4 ± 1.0

Table 1: The OA is normalized on the total number of samples; TN and TP are normalized on the true values, i.e. on the rows. The classification precision improves by using the *synthetic* features for the algorithms training.

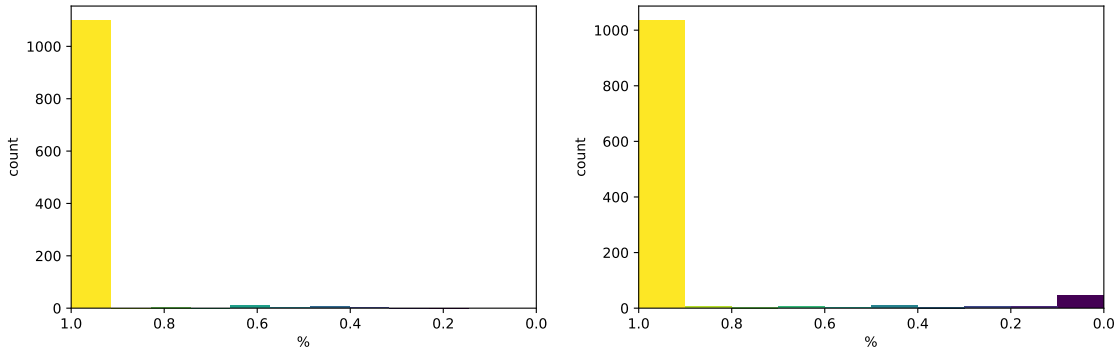

Figure 3: Classification histogram of *Fermi*-LAT unIDs with a NN trained on standardized data with 2F (left panel) and 4F (right panel). The color scale represents the probability for being classified as astro, i.e. yellow bar indicates a 100% classification as astrophysical source (Astro) and magenta bar indicates a 100% classification as DM target (0% as Astro).

4. Preliminary results

We have classified 1125 unIDs with the trained NN. Fig. 3 shows the classification results for the two-feature (left panel) and four-feature classification (right panel). Thus, we find out that:

- Classification with two features: 1116 unIDs have a probability $\geq 50\%$ to be astrophysical sources; 3 unIDs have a probability $\geq 60\%$ to be DM target;
- Classification with four features: 1053 unIDs have a probability $\geq 50\%$ to be astrophysical sources; 33 unIDs have a probability $\geq 99\%$ to be DM target (45 unIDs have a probability $\geq 90\%$ to be DM targets).

Indeed, by introducing the *synthetic* features (i.e. systematic uncertainty) a number of unIDs is reconsidered as prospective DM source.

5. Conclusions

In these proceedings we present first preliminary results of our search for DM targets among Fermi-LAT unIDs with NN. The algorithm is trained on a parameter space of both experimental and theoretical sample, the latter being enriched by the introduction of *synthetic* features, which indirectly allow us to include experimental systematics in the DM set. Further efforts will be focused to study different strategy in order to include the same uncertainty within different classification algorithms, e.g. Gaussian Processes with noisy input [7].

Acknowledgments

The work of VG, JCB, MASC and BZ was supported by the Spanish Agencia Estatal de Investigación through the grants PGC2018-095161-B-I00 and IFT Centro de Excelencia Severo

[1]Preliminary results obtained with a balanced data set, by including all σ_{TS} and standardized data.

Ochoa SEV-2016-0597, the Atracción de Talento contract no. 2016-T1/TIC-1542 granted by the Comunidad de Madrid in Spain, and the MultiDark Consolider Network FPA2017-90566-REDC. VG's contribution to this work has been supported by *Juan de la Cierva-Formación* FJCI-2016-29213 and *Juan de la Cierva-Incorporación* IJC2019-040315-I grants. BZ has been further supported by the Programa Atracci´on de Talento de la Comunidad de Madrid under grant n. 2017-T2/TIC-5455, from the Comunidad de Madrid/UAM "Proyecto de J´ovenes Investigadores" grant n. SI1/PJI/2019-00294, from Spanish "Proyectos de I+D de Generacion de Conocimiento" via grant PGC2018-096646-A-I00. BZ finally acknowledge the support from Generalitat Valenciana through the plan GenT program (CIDEGENT/2020/055).

A. Classification algorithm parameters in scikit-learn

In order to ensure the reproducibility of our results, in this appendix we report the parameters adopted in scikit-learn for this study.

Logistic Regression

We use the *sklearn.linear_model.LogisticRegression* tool of [15] with the (solver='lbfg', random_ state=0) and other default options. Note that regularization is applied by default.

Artificial Neural Network

We use the *sklearn.neural_ network.MLPClassifier* [15]. Our entries are: (solver='adam', alpha=0.0, batch_ size=120, hidden_ layer_ sizes=(41,), learning_ rate_ init=0.015, max_ iter=1000, random_ state=0, activation='relu') and other default options.

References

[1] M. Feickert and B. Nachman, *A Living Review of Machine Learning for Particle Physics*, 2102.02770.

[2] W. Atwood, A. Albert, L. Baldini, M. Tinivella, J. Bregeon, M. Pesce-Rollins et al., *Pass 8: Toward the full realization of the fermi-lat scientific potential*, 2013.

[3] Fermi-LAT collaboration, *Fermi Large Area Telescope Fourth Source Catalog*, *Astrophys. J. Suppl.* **247** (2020) 33 [1902.10045].

[4] A. Bhat and D. Malyshev, *Machine learning methods for constructing probabilistic Fermi-LAT catalogs*, 2102.07642.

[5] C.Y. Hui, J. Lee, K.L. Li, S. Kim, K. Oh, S. Luo et al., *Searches for pulsar-like candidates from unidentified objects in the Third Catalog of Hard Fermi-LAT Sources with machine learning techniques*, *Mon. Not. Roy. Astron. Soc.* **495** (2020) 1093 [2004.10945].

[6] M. Kovačević, G. Chiaro, S. Cutini and G. Tosti, *Optimizing neural network techniques in classifying Fermi-LAT gamma-ray sources*, *Mon. Not. Roy. Astron. Soc.* **490** (2019) 4770 [1911.02948].

[7] C. Villacampa-Calvo, B. Zaldivar, E.C. Garrido-Merchán and D. Hernández-Lobato, *Multi-class Gaussian Process Classification with Noisy Inputs*, *arXiv e-prints* (2020) arXiv:2001.10523 [2001.10523].

[8] S. Germani, G. Tosti, . Lubrano, S. Cutini, I. Mereu and A. Berretta, *Artificial Neural Network Classification of 4FGL Sources*, 2106.08222.

[9] G. Bertone and D. Merritt, *Dark matter dynamics and indirect detection*, *Mod. Phys. Lett.* **A20** (2005) 1021 [astro-ph/0504422].

[10] G. Bertone, N. Bozorgnia, J.S. Kim, S. Liem, C. McCabe, S. Otten et al., *Identifying WIMP dark matter from particle and astroparticle data*, *JCAP* **03** (2018) 026 [1712.04793].

[11] J. Coronado-Blázquez, M.A. Sánchez-Conde, M. Di Mauro, A. Aguirre-Santaella, I. Ciucă, A. Domínguez et al., *Spectral and spatial analysis of the dark matter subhalo candidates among Fermi Large Area Telescope unidentified sources*, *JCAP* **11** (2019) 045 [1910.14429].

[12] J.A.R. Cembranos, A. de la Cruz-Dombriz, V. Gammaldi, R.A. Lineros and A.L. Maroto, *Reliability of Monte Carlo event generators for gamma ray dark matter searches*, *JHEP* **09** (2013) 077 [1305.2124].

[13] M. Cirelli, G. Corcella, A. Hektor, G. Hutsi, M. Kadastik, P. Panci et al., *PPPC 4 DM ID: A Poor Particle Physicist Cookbook for Dark Matter Indirect Detection*, *JCAP* **03** (2011) 051 [1012.4515].

[14] M. Zięba, S.K. Tomczak and J.M. Tomczak, *Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction*, *Expert Systems with Applications* **58** (2016) 93.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel et al., *Scikit-learn: Machine learning in Python*, *Journal of Machine Learning Research* **12** (2011) 2825.

PoS(ICRC2021)493