Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **520,** 1348–1361 (2023)
Advance Access publication 2023 January 13

https://doi.org/10.1093/mnras/stad066

A search for dark matter among Fermi-LAT unidentified sources with systematic features in machine learning

V. Gammaldi ,[1,2]★ B. Zaldívar,[3]★ M. A. Sánchez-Conde[1,2]★ and J. Coronado-Blázquez[1,2,4]

[1] *Departamento de Física Teórica, Universidad Autónoma de Madrid, Cantoblanco, E-28049 Madrid, Spain*
[2] *Instituto de Física Teórica, IFT-UAM/CSIC, Cantoblanco, E-28049 Madrid, Spain*
[3] *Institute of Corpuscular Physics (IFIC), University of Valencia and CSIC, Calle Catedrático José Beltrán 2, E-46980 Paterna, Spain*
[4] *Telefónica Tech IoT and Big Data, Ronda de la Comunicación s/n, E-28050 Madrid, Spain*

Accepted 2023 January 6. Received 2022 December 12; in original form 2022 July 29

ABSTRACT

Around one-third of the point-like sources in the Fermi-LAT catalogues remain as unidentified sources (unIDs) today. Indeed, these unIDs lack a clear, univocal association with a known astrophysical source. If 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 from WIMPs annihilation. We propose a new approach to solve the standard, machine learning (ML) binary classification problem of disentangling prospective DM sources (simulated data) from astrophysical sources (observed data) among the unIDs of the 4FGL Fermi-LAT catalogue. We artificially build two *systematic* features for the DM data which are originally inherent to observed data: the detection significance and the uncertainty on the spectral curvature. We do it by sampling from the observed population of unIDs, assuming that the DM distributions would, if any, follow the latter. We consider different ML models: Logistic Regression, Neural Network (NN), Naive Bayes, and Gaussian Process, out of which the best, in terms of classification accuracy, is the NN, achieving around 93.3 per cent ± 0.7 per cent performance. Other ML evaluation parameters, such as the True Negative and True Positive rates, are discussed in our work. Applying the NN to the unIDs sample, we find that the degeneracy between some astrophysical and DM sources can be partially solved within this methodology. None the less, we conclude that there are no DM source candidates among the pool of 4FGL Fermi-LAT unIDs.

Key words: astroparticle physics – methods: data analysis – methods: observational – methods: statistical – dark matter – gamma-rays: general.

1 INTRODUCTION

Astrophysical and cosmological evidence suggests that non-baryonic cold DM constitutes 84 per cent of the matter density of the Universe (Ade et al. 2016; Aghanim et al. 2020). Although the nature of DM is still unknown, weakly interacting massive particles (WIMPs) are popular and well-motivated DM candidates, among others. In particular, WIMPs are one of the most popular types of DM candidates in the context of DM searches. The WIMP paradigm invokes the same thermal decoupling, which is enormously successful at making detailed predictions for many observables in the early Universe, including the abundances of light elements and the CMB (Peebles et al. 1991). Indeed, it is somewhat natural to invoke a similar paradigm to infer the abundance of DM as a thermal relic from the early Universe. The reason is that, in order to fill all the DM content that we observe in the universe, WIMPs should have sizeable interactions with the Standar Model (SM) sector, thus ensuring a rich phenomenology while still having a relevant part of their parameter space allowed from all the available experimental data. In the so-called indirect detection searches, and in particular those relying on gamma-ray measurements in the Fermi-LAT's energy range, WIMPs are the natural candidates to consider given their expected masses. In fact, WIMPs are predicted to annihilate or decay into SM particles, whose decay and hadronization processes would produce secondary particles, such as cosmic rays, neutrinos, and gamma-rays (Buckley & Hooper 2010; Zechlin et al. 2011; Belikov, Buckley & Hooper 2012; Zechlin & Horns 2012; Berlin & Hooper 2013; Bertoni, Hooper & Linden 2015, 2016; Calore et al. 2016; Schoonenberg et al. 2016; Hooper & Witte 2017). The flux of secondary particles may be observed in ground-based or satellite observatories, laying the groundwork for the indirect searches for DM. Only an agreement of several hints in the observed flux of different messengers – i.e. the multimessenger detection – would result in a competitive claim of the indirect detection of DM (Bergström 2013; Gammaldi 2019). This includes the issue of disentangling the DM signal from the emission of well-known astrophysical sources or diffuse astrophysical background. Indeed, DM-dominated systems – e.g. dwarf galaxies, galaxy clusters as well as the Galactic Centre – are benchmark targets for indirect searches for DM (see e.g. Charles et al. 2016; Conrad & Reimer 2017; Gammaldi et al. 2021 and refs therein). Among others, gamma-rays are considered to be the golden messenger: they are (very-) high-energy neutral particles travelling practically undeflected along straight paths in the local Universe.

★ E-mail: viviana.gammaldi@uam.es (VG); b.zaldivar.m@csic.es (BZ); miguel.sanchezconde@uam.es (MAS-C)

The Large Area Telescope (LAT) on-board the NASA *Fermi* satellite (*Fermi*-LAT) (Atwood et al. 2013) has collected more than 13 yr of gamma-ray data of the full sky. Still in operation, *Fermi*-LAT is a pair conversion telescope capable to observe gamma-ray photons from energies ∼20 MeV up to ∼ TeV. Several point-source catalogues have been released and contain thousands of gamma-ray objects, many of them previously unknown (The Fermi-LAT Collaboration 2015, 2016, 2017). Interestingly, around one-third of the point-like gamma-ray sources in the 4FGL *Fermi*-LAT catalogue (Abdollahi et al. 2020) as well as in other gamma-ray ground-based telescopes (see e.g. Ahnen et al. 2019) remain as unidentified (unIDs) today. These unIDs lack a clear, univocal association with a known astrophysical source.

In the last few years, machine learning (ML) techniques have been applied to many different fields of astrophysics and cosmology; e.g. applied to the so-called Galactic Centre Excess (Caron et al. 2018), to the search for dark matter in dwarf galaxies (Calore, Serpico & Zaldivar 2018; Alvarez et al. 2020), as well as classification algorithms that have been applied to the *Fermi*-LAT catalogues (see e.g. Mirabal et al. 2016; Bartels & Edwards 2019; Kovačević et al. 2019; Hui et al. 2020; Villacampa-Calvo et al. 2020; Germani et al. 2021; Bhat & Malyshev 2022; and references therein). The latter works have been focused on classifying unIDs as different types of known astrophysical sources (e.g. Active Galactic Nuclei, pulsars, blazars). None the less, if DM is made of WIMPs, there is also the exciting possibility that some of these unIDs may actually be DM sources, emitting gamma-rays by WIMPs annihilation (Bertone & Merritt 2005). In fact, the nature of DM still represents an open question in physics and cosmology, and many efforts have been devoted to understand its nature via the application of novel ML techniques in several related fields[1] (e.g. Agarwal, Davé & Bassett 2018; Bertone et al. 2018; Morice-Atkinson, Hoyle & Bacon 2018; Feickert & Nachman 2021; Spencer et al. 2021; Ullmo, Decelle & Aghanim 2021; Bazarov et al. 2022; Holwerda et al. 2022).

Around one-third of the point-like sources in the Fermi-LAT catalogues remain as unidentified sources (unIDs) today. Indeed, these unIDs lack a clear, univocal association with a known astrophysical source. If dark matter (DM) is composed of WIMPs, there is the exciting possibility that some of these unIDs may actually be DM sources, emitting gamma-rays from WIMPs annihilation. We propose a new approach to solve the standard, machine learning (ML) binary classification problem of disentangling prospective DM sources (simulated data) from astrophysical sources (observed data) among the unIDs of the 4FGL Fermi-LAT catalogue. Concretely, we artificially build two systematic features for the DM data which are originally inherent to observed data: the detection significance and the uncertainty on the spectral curvature. We do it by sampling from the observed population of unIDs, assuming that the DM distributions would, if any, follow the latter. We consider different ML models: Logistic Regression, Neural Network (NN), Naive Bayes and Gaussian Process, out of which the best, in terms of classification accuracy, is the NN, achieving around 93.3. In this work, we propose a new approach to solve the binary classification problem of disentangling prospective DM-source candidates from astrophysical sources among the unIDs in the 4FGL *Fermi*-LAT catalogue. We work on the derived parameter space defined by the energy-peak E_{peak} and curvature β of the gamma-ray spectra of source in the catalogue: the so-called *Fermi*-LAT β-plot (Coronado-Blázquez et al. 2019a). The observational β-plot – composed of both identified and unidentified gamma-ray sources – will be here enriched by theoretically based DM parameters and (hereafter the so-called ‘DM-β’ plot).

Many works have pointed out the spectral confusion between pulsars and DM annihilation signals in gamma-rays (e.g. The Fermi-LAT Collaboration 2012; Mirabal 2013; Mirabal et al. 2016), which is especially relevant when considering light, $\mathcal{O}$ (10GeV) WIMPs, and hadronic annihilation channels such as $b\bar{b}$. Indeed, such a degeneracy is pictured as an overlapping region in the $E_{\mathrm{peak}} - \beta$ plane. None the less, in our work we will show that WIMP candidates cover a broader region in this parameter space. Hereafter, we refer to the parameters of such a plot as *features*, by using the benchmark ML nomenclature. Because of the present degeneracy in the $E_{\mathrm{peak}} - \beta$ plane, we introduce two *systematic* features for the DM sample, motivated by the systematic uncertainty of the *Fermi*-LAT detector, which would affect the detection of any DM source. This allow us to train the classification algorithms with four features (4F) instead of 2F. Furthermore, we also discuss the possibility to adopt a three-feature setup, by including the relative uncertainty on β (β_{rel}) via both a sampled Gaussian distribution of the uncertainty itself (3F-A) and in the statistical model (3F-B).

We consider four classification algorithms, namely, Logistic Regression (LR), Neural Network (NN), Naïve Bayes (NB), and Gaussian Process (GP). The LR and NN algorithms are built in the `scikit-learn` library for data analysis with ML in python (Pedregosa et al. 2011), while we implement our own python codes for NB and GP models, the latter using `tensorflow v1`, the open-source library for automatic differentiation and ML applications (Developers 2022).

These four classification models have been selected according to their different advantages and capabilities[2]: LR is arguably the simplest model (it gives linear decision boundaries among the classes of points) and consequently it is highly explainable, even though it requires numerical optimization. NN on the other hand is, a priori, arbitrarily expressive while at the same time being optimized very efficiently, the reason for which it is one of the most popular ML models for problems in a wide range of domains. NB is a model giving a priori a higher expressive power than LR (it can give non-linear decision boundaries) while requiring analytical optimization. Finally, a GP classifier (Rasmussen & Williams 2006) offers as well a high expressivity with the added value of being a Bayesian model, allowing us to report prediction uncertainties.

This paper is organized as follows: in Section 2 we introduce the data. In Section 3 we introduce the *systematic* features that will be used by our algorithms. In Section 4 we introduce the methodology, with the classification algorithms and feature setups. In Section 5 we compute the ‘DM-versus-astrophysics’ classification accuracy for the selected algorithms under different setups. In Section 6 we provide the results of the classification of unIDs with our best classifier from the previous exercise, before concluding in Section 7.

2 EXPERIMENTAL AND THEORETICAL DATA

The recent 4FGL *Fermi*-LAT catalogue (Abdollahi et al. 2020) is a collection of sources with associated gamma-ray spectra, containing important information about their nature. Somehow surprisingly, an important fraction of objects in the *Fermi*-LAT catalogues, ca. 1/3 of the total, remain as unIDs, i.e. objects lacking a clear single

[1] See e.g. darkmachines.org.

[2] See Bishop (2006), one of the standard reference books for Machine Learning.

association to a known object identified at other wavelengths, or to a well-known spectral type emitting only in gamma-rays, e.g. certain pulsars. Among other prospective sources of gamma-rays from DM annihilation events, dark satellites, or subhaloes in the Milky Way, with no optical counterparts, are the preferred candidates, as they are expected to exist in high number according to standard cosmology and they would not be massive enough to retain gas/stars, this way being pristine DM annihilating sources free of gamma-ray astrophysical backgrounds. Many authors have already investigated DM subhaloes as prospective targets for indirect DM detection (Buckley & Hooper 2010; Zechlin et al. 2011; Belikov et al. 2012; The Fermi-LAT Collaboration 2012; Zechlin & Horns 2012; Berlin & Hooper 2013; Belotsky, Kirillov & Khlopov 2014; Bertoni et al. 2015, 2016; Calore et al. 2016; Schoonenberg et al. 2016; Hooper & Witte 2017; Coronado-Blázquez et al. 2019a, b; Coronado-Blázquez et al. 2022).

The *Fermi*-LAT 4FGL catalogue (Abdollahi et al. 2020) adopted in this work, is the result of 8 yr of telescope operation. It covers the 50 MeV–1 TeV energy range, and reports the detection of over 5000 gamma-ray sources, almost doubling the previous 3FGL, and using the latest instrumental response functions (IRFs) and Pass 8 events (Atwood et al. 2013), which optimize the instrument capacities, as well as an updated Galactic diffuse emission model. In particular, we are interested in one of the parametrizations of the gamma-ray spectrum used in the 4FGL, known as the Log-Parabola (LP):

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

where N_0 is the gamma-ray flux normalization, E_0 the pivot energy, α the gamma-ray spectral index, and β the curvature. Note that this parametric form is reduced to a simple power law in the case of $\beta = 0$. From this expression we can extract a useful parameter: the peak energy, E_{peak}, i.e. the energy at which the energy spectrum ($E^2\mathrm{d}N/\mathrm{d}E$) is maximum, by performing the consequent derivative, obtaining $E_{\mathrm{peak}} = E_0 \cdot e^{\frac{2-\alpha}{2\beta}}$, which represents a signature of different kind of emitting sources. In this work, we do not perform the spectral analysis of the 4FGL catalogue sources by ourselves. Instead, we use the parameters of the LP fit for each source published by the Fermi-LAT collaboration.

Similarly, we can predict the gamma-ray DM spectrum by means of Monte Carlo event generator softwares (see e.g. Cirelli et al. 2011; Cembranos et al. 2013). In fact, WIMPs annihilate in different SM channels, whose 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. In Coronado-Blázquez et al. (2019a), the authors introduced the DM in the $\beta - E_{\mathrm{peak}}$ parameter space (i.e. the β-plot) by fitting the DM gamma-ray spectrum, given by Cirelli et al. (2011), with the same LP functional form (equation 1). While in Coronado-Blázquez et al. (2019a) only pure annihilation channels ($B_r = 1$) were studied, we now consider more general two-channel linear combinations, of the form

$$\frac{\mathrm{d}N}{\mathrm{d}E} = B_r \left(\frac{\mathrm{d}N}{\mathrm{d}E}\right)_{C_1} + (1 - B_r)\left(\frac{\mathrm{d}N}{\mathrm{d}E}\right)_{C_2}, \tag{2}$$

where C_1 and C_2 are the two considered channels. We perform all possible combinations considering 10 branching ratios from 0 to 1 with a 0.1 step, for the annihilation channels $b\bar{b}$, $c\bar{c}$, $t\bar{t}$, $\tau^+\tau^-$, e^+e^-, $\mu^+\mu^-$, W^+W^-, Z^0Z^0, and hh, and masses from 5 GeV[3] to 10 TeV.[4] As we are agnostic to the underlying particle physics model that generates the annihilation, we consider all points as a 'DM cloud' – therefore being able to distinguish only between the astrophysical and DM scenarios, which is the ultimate goal of this paper. We generate a convenient number of DM points randomly distributed within the boundaries of the DM parameter space. The 'DM-β'-plot is shown in Fig. 1. In this plot, the orange points are astrophysical gamma-ray sources, the red points are detected unIDs and the magenta points are the DM sample. The overlap between DM and astrophysical sources is for light WIMPs and pulsars mainly, and especially in the case of hadronic channels such as $b\bar{b}$ and $c\bar{c}$, as expected (The Fermi-LAT Collaboration 2012; Mirabal 2013; Mirabal et al. 2016). None the less, a good portion of the region of the parameter space where the DM resides is radically different from the one where astrophysical sources lie.

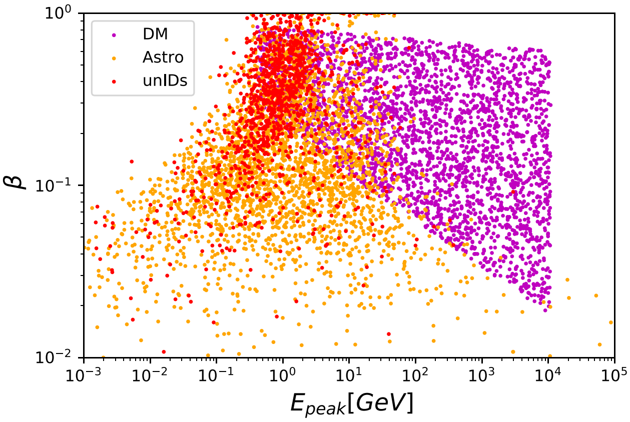

Figure 1. The 'DM-β plot', which includes information about the gamma-ray spectra of well-known astrophysical gamma-ray sources (orange points), unIDs sources (red points), and theoretical WIMP DM sources data set (magenta points).

3 DARK MATTER SYSTEMATIC FEATURES

In the previous section, we have summarized and generalized the methodology of Coronado-Blázquez et al. (2019a) in order to introduce the WIMPs candidates in the β-plot parameter space, which allows us to train ML algorithms in order to distinguish and classify prospective DM-source candidates from astrophysical sources, only based on their gamma-ray spectra.

None the less, such a description of the DM sample with only the two features of the β-plot, represents a limitation in the framework of ML. In fact, the collection of the unIDs sources we aim to classify includes a plethora of information – in terms of data or number of features – that are not considered in such a phenomenological DM data set. Among other observational features that are not yet available for the DM sample, we will consider, on the one hand, the

[3] In some cases the lower mass is bounded by the mass of the particle itself, namely for the annihilation channels with $W^{\pm}$ ($m_{W^{\pm}} = 80$ GeV), Z^0 ($m_{Z^0} = 91$ GeV), h ($m_h = 125$ GeV) and $t/\bar{t}$ ($m_{t/\bar{t}} = 173$ GeV).

[4] Although the spectra from Cirelli et al. (2011) go to masses up to 100 TeV, the model-independent electroweak corrections used in these calculations are computed at leading order, while masses larger than ~10 TeV, especially in leptonic channels, lack higher order electroweak corrections not included in the tables, which may be relevant (Ciafaloni et al. 2011; Cirelli et al. 2011). In any case, the LAT sensitivity quickly degrades at energies $\gtrsim 300$ GeV.

experimental systematic uncertainty $\beta_{\rm rel} = \varepsilon_\beta/\beta$, of the curvature parameter β, and on the other hand the detection significance of the source, σ_d. Both quantities are of course inherent to both the identified sources and the unIDs. We explain below our procedure to artificially build such quantities for the DM sample.

3.1 Detection significance

First of all, it is phenomenologically interesting to note the different spread of the astrophysical classes in the β-plot. In Fig. 2 we show how the overlap between different astrophysical sources decreases by changing the cut applied on the detection significance, namely $\sigma_{\rm d} \geq 4,\ 10,\ 50$: the larger the significance, the smaller the overlap.

Generally speaking, the detection significance strictly depends on the data analysis. To analyse LAT data, the collaboration tools construct the likelihood that is applicable to the LAT data, and then use this likelihood to find the best-fitting model parameters. These parameters include the description of a source's spectrum, its position, and even whether it exists. Once that a template model of all the other sources in the source region is provided, the Test Statistic (TS) for adding an additional source at each grid point is calculated. The resulting significance is $\sim(TS)^{1/2}\sigma_d$, and thus TS = 16–25 equivalent to 4–5σ_d, is required for claiming the detection of any source in the 4FGL Fermi-LAT catalogue adopted in this work. The new source is characterized by a source intensity and spectral index.[5] Hereafter, we will use the so-defined detection significance σ_d as a *systematic* feature of our classification problem. Note that, the DM data set has been created based on the WIMP phenomenology and the procedure outlined in Section 2, and thus, it obviously lacks a detection significance, which is – by definition – an observational feature. In order to exploit the additional information coming from the distribution of the detection significance of the detected astrophysical sources, our idea is to build this variable as a fictitious feature of the DM class. The issue is not straightforward: in fact, the detection significance $\sigma_{\rm d}$ for the prospective DM sources would ultimately depend on many aspects, e.g. the WIMP mass, the SM annihilation channel, the Monte Carlo event generator software (Cembranos et al. 2013), the distance of the sources, the amount of DM in the source, as well as other hypotheses on the DM particle (see e.g. Visinelli 2018). If several DM subhaloes were discovered, this class of DM sources would follow its own $\sigma_{\rm d}$ distribution (see e.g. Section 3 of Gammaldi et al. 2021).

As first hypothesis, we can assume that all the DM-source candidates are among the observed unIDs. We can therefore sample the unIDs $\sigma_{\rm d}$ distribution to generate mock data for DM with a random noise (Fig. 3), such that the distribution is statistically the same but a single DM point in the DM-β plot is assigned a random $\sigma_{\rm d}$. In this way, we associate to the theoretical DM sample a *systematic* feature, which reflects systematics related to the adopted instrument, as shown in Fig. 7, first upper panel.

3.2 Uncertainty on β

The next step of this analysis relies on the intuition that higher values of the detection significance $\sigma_{\rm d}$ correspond to better signal-to-noise ratio, i.e. to higher quality source spectra. When dealing with actual sources, the *Fermi*-LAT standard analysis pipeline bins the spectral energy distribution (SED) and fits it with the corresponding parametric form (here, a log-parabola). The uncertainty in each bin

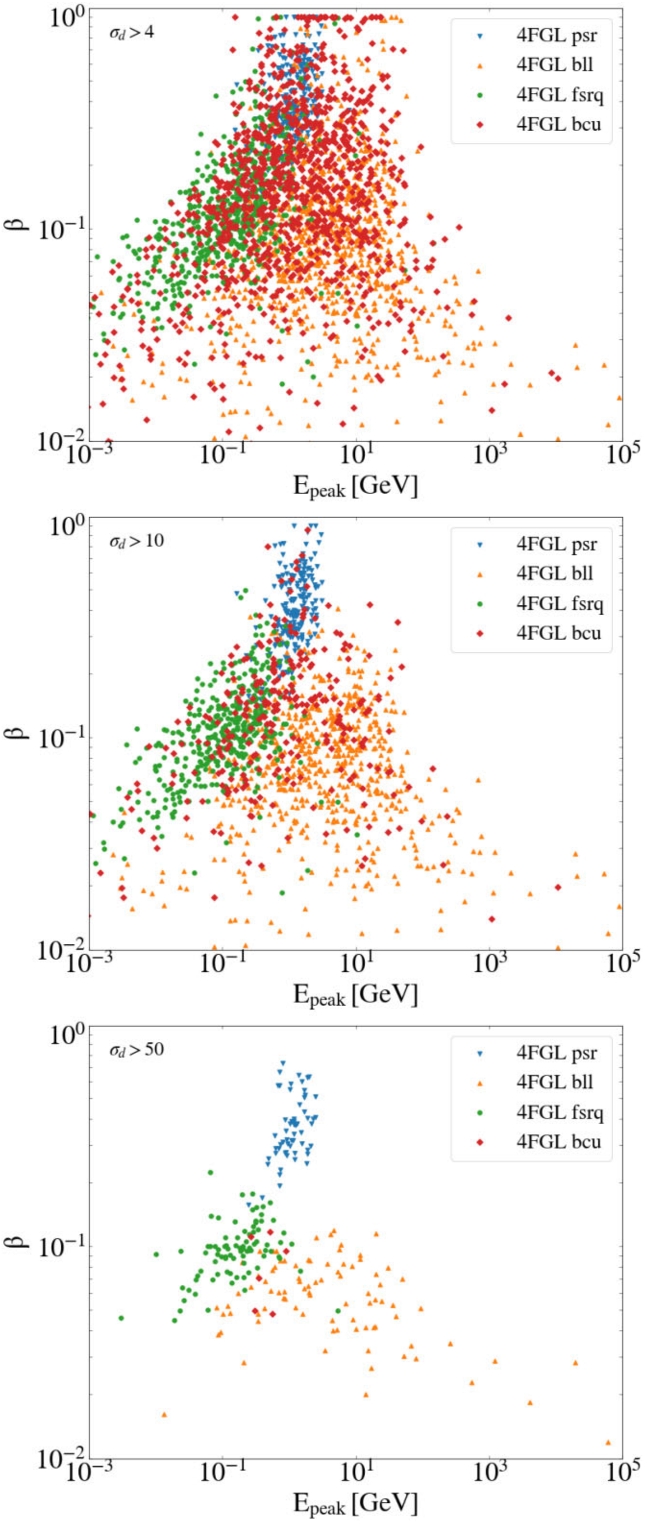

Figure 2. The same as Fig. 1 for 4FGL identified sources only and different cuts in detection significance $\sigma_{\rm d}$. Top panel: $\sigma_d > 4$ (all sources). Middle panel: $\sigma_d > 10$. Bottom panel: $\sigma_d > 50$. Note the better separability of the classes as the cut is more stringent, at the cost of reducing the sample.

[5] In a first approximation, the spectrum is assumed to be a power law.

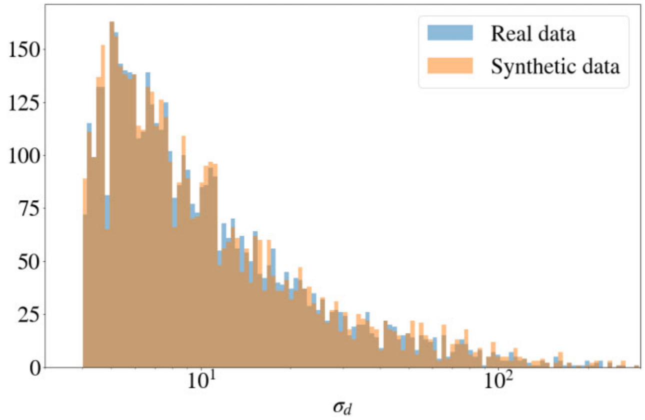

Figure 3. Detection significance (σ_{d}) distribution for the real unIDs data (blue) and systematic sampling for DM (orange), with a random noise similar to the one seen in the unIDs sample. The brown colour only reflects the overlap of the previous two distributions.

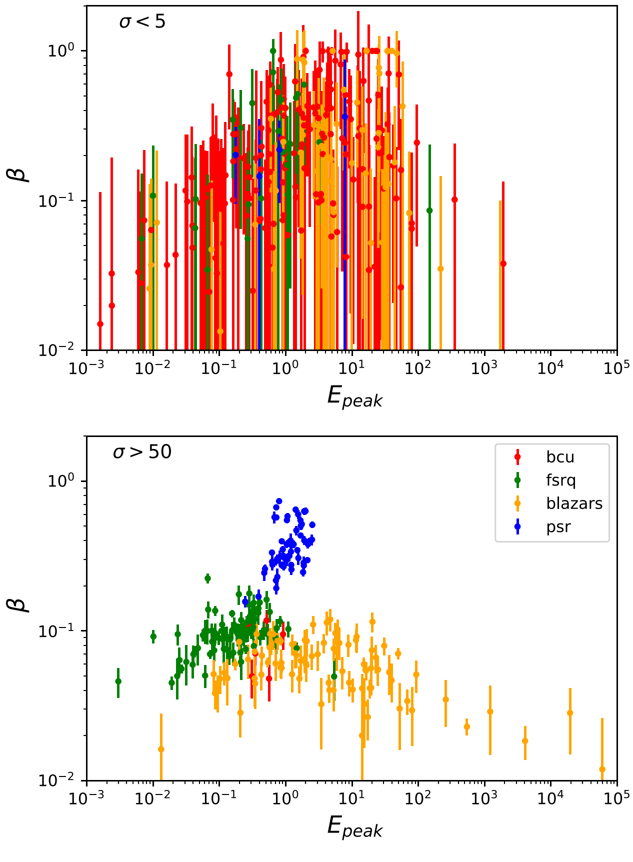

Figure 4. Same as Fig. 2, by including the uncertainty on β for astrophysical data with $\sigma_d < 5$ (upper panel) and $\sigma_d > 50$ (lower panel). Let us stress as a lower detection significance corresponds to a worse characterization of the spectrum. Indeed, the classification is confused for data of lower σ_d and improves for higher values of σ_d also by eye. The colour code in the legend is the same for both panels.

will translate into an error in the parameters of the model. As a consequence, we expect a lower detection significance to correspond to a worse characterized spectrum. This translates into a higher uncertainty in the estimation of the spectral parameters (E_{peak}, β). We show qualitatively this property in Fig. 4 for astrophysical data with $\sigma_d < 5$ (upper panel) and $\sigma_d > 50$ (lower panel). The correlation between the relative error $\beta_{\mathrm{rel}} = \epsilon_\beta/\beta$ and the σ_{d} is shown in the upper panel of Fig. 5 for the unIDs population: clearly, the relative error $\beta_{\mathrm{rel}} = \epsilon_\beta/\beta$ decreases by increasing the detection significance σ_{d},[6] although they are not completely correlated. In the lower panel of Fig. 5 we show the correlation between the relative error β_{rel} and E_{peak}: in this case the correlation is slightly visible and it could be associated to the sensitivity of the instrument to different energy bins. In other words, there are more detected sources (with lower σ_{d}) in the energy range where LAT is more sensitive (~1 GeV), as expected. While partial correlations may exist among all the features considered in our analysis, a generic property of typical ML models is that they take them into account automatically (Bishop 2006). This is exactly the kind of information that we aim to implicitly include via the analysis of these data within a ML approach instead of benchmark analyses.

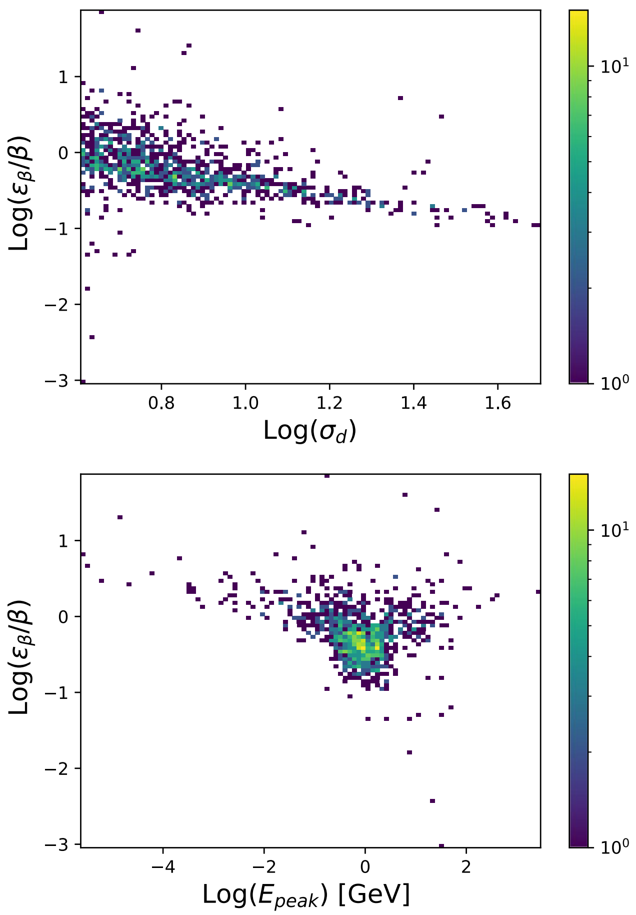

Figure 5. Upper panel: Relative error $\beta_{\mathrm{rel}} = \varepsilon_\beta/\beta$ versus the detection significance σ_{TS} for the 4FGL unIDs. Lower panel: Relative error, ε_β/β, versus the E_{peak} for the 4FGL unIDs. Although there are sources with $E_{\mathrm{peak}} < 0.3$ GeV, no DM point lie below this value, as the lightest WIMP mass considered is 5 GeV and the softest channel is $b\bar{b}$, which roughly peaks at $E_{\mathrm{peak}} \sim m_\chi/20$).

As in the σ_{d} case, the DM data lack an observational β_{rel}. Although a purely theoretical β_{rel} is given by the LP fitting of the simulated gamma-ray spectra expected by DM annihilation events (Cirelli et al. 2011), we verified that such a theoretical error is below a few per cent,

[6] Probably due to a Poissonian statistical contribution to the total uncertainty.

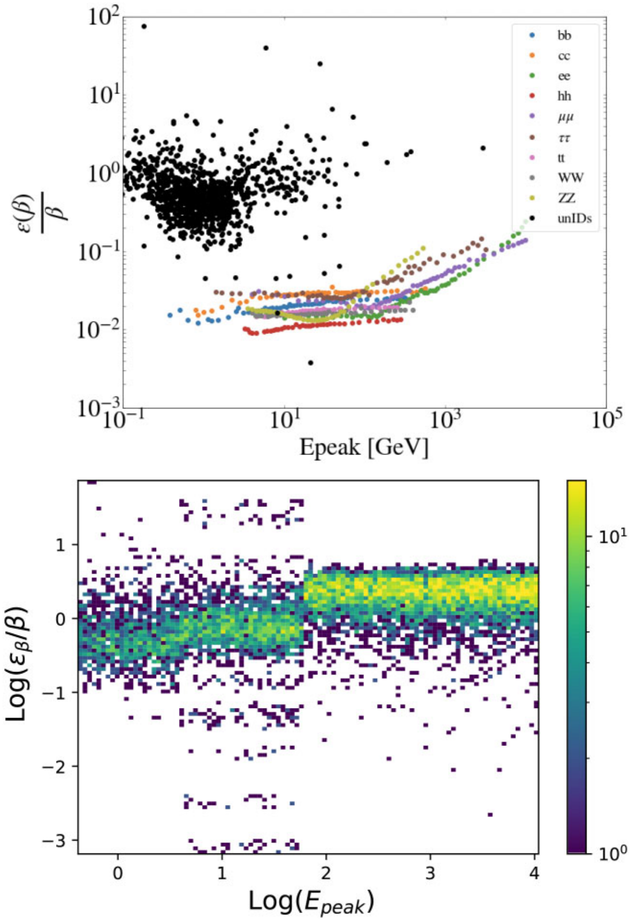

Figure 6. Upper panel: experimental versus theoretical $\beta_{\rm rel}$. Lower panel: relative error, $\beta_{\rm rel} = \varepsilon_\beta/\beta$, versus the $E_{\rm peak}$ for the DM data, sampled from the 4FGL unIDs. Three regimes are considered, $E_{\rm peak} < 4$ GeV, $4 < E_{\rm peak} < 60$ GeV, and $E_{\rm peak} > 60$ GeV.

and can be neglected with respect to the much larger systematic $\beta_{\rm rel}$ discussed so far. Such comparison is shown in the upper panel of Fig. 6. The adopted distribution is shown in the lower panel of Fig. 6 and second panel of Fig. 7.

3.3 $\beta_{\rm rel}$ as systematic feature

Analogously to the $\sigma_{\rm d}$ feature, we can sample the distribution of $\beta_{\rm rel}$ for the unIDs population to associate uncertainties on β to the DM data. Here we consider a 2D sampling space, as the ε_β depends also on the $E_{\rm peak}$. Indeed, from Fig. 5 (lower panel), one can see that there is a cluster at $E_{\rm peak} \sim 0.5$–2 GeV, while for energies above ~60 GeV the errors tend to be larger as the statistics of unIDs decreases. This is due to the LAT sensitivity, which reaches its maximum at 1–2 GeV.

In order to assign the DM $\beta_{\rm rel}$ systematic values, we will divide the distribution in three bins, $E_{\rm peak} < 4$ GeV, $4 < E_{\rm peak} < 60$ GeV, and $E_{\rm peak} > 60$ GeV. According to Fig. 5, these boundaries approximately reflect three different regimes in the data, with a cluster of objects in the first one, a more spread distribution in the second and a third one where no source with $\varepsilon_\beta/\beta < 1$ is found. Sampling directly the unbinned distribution would lead to underestimated errors for the highest $E_{\rm peak}$ values. As the points with $E_{\rm peak} > 60$ GeV are very scarce (just 7), we introduce a random Gaussian noise to avoid discreteness in the distribution. For consistency, we do the same for the other two bins. The result of this sampling is shown in the lower panel of Fig. 6. The last bin is the most populated one, as it is the one which contains more DM points (mostly due to hard channels such as e^+e^- and $\mu^+\mu^-$, which peak at $E_{\rm peak} = m_\chi$). The pronounced step visible in the figure at $E_{\rm peak} = 60$ GeV is simply caused by the binning choice, and can also be seen in the original unIDs distribution of Fig. 5.

With these two samplings of $\sigma_{\rm d}$ and $\beta_{\rm rel}$, we have generated two equivalent data sets consisting on $\{E_{\rm peak}, \beta, \beta_{\rm rel}, \sigma_d\}$, both for the 4FGL catalogue and DM. The *systematic* distribution for σ_d and $\beta_{\rm rel}$ created for the theoretical DM sample (magenta histograms) are shown in the two lower panels of Fig. 7 – as well as the distributions for the observed data, i.e. astrophysical sources (orange histograms) and the unIDs (red histograms). Note also that, although we adopt in our analysis the LP parameters given by the Fermi-LAT collaboration for the full catalogue of detected sources, we take into account the possibility that any Fermi-LAT source could be not well fitted with an LP by including the $\beta_{\rm rel}$ uncertainty in the 4F analysis. in Appendix A and Fig. A1 we show an example of the β-plot with the astrophysical and DM data set including the systematic uncertainty on β.

4 METHODOLOGY

4.1 Classification algorithms

We interpret the problem as a standard binary classification task in ML. Data consists of $\mathcal{D} = \{\mathbf{x}_i, t_i\}$, being $i = 1,..., N$, where $\mathbf{x} = \{E_{\rm peak}, \beta, \beta_{\rm rel}, \sigma_d\}$ is the multivariate input and a label $t_i = \{0, 1\}$, corresponding to astro or DM class, respectively.

We study the performance of several ML models, coming from different approaches and having different levels of expressiveness. They are briefly specified next:

(i) Probabilistic discriminative models. In order to estimate the expected value y_i of the label t_i, we start with the simplest classifier: the Logistic Regression (LR) model. Secondly, we use a fully connected feed-forward neural network (NN) with one hidden layer. See Appendix C for further technical details of the implementation. Both models aim at estimating the probability $p(C_k|\mathbf{x})$ of class C_k given the input $\mathbf{x}$, which will depend on some parameters to be optimized.

(ii) Generative model. We also consider the Naïve Bayes (NB) classifier, where the likelihood $p(\mathbf{x}_i|C_k)$ of point i given a class C_k is taken as a multivariate Gaussian distribution, with diagonal covariance matrix. This is a common benchmark model, since even if not requiring numerical optimization, it is typically giving reasonably good results also in real-world data sets. We have used our own `python` implementation of this model.

(iii) Non-parametric model. Finally, we consider a specific Gaussian Process classifier,[7] namely Noisy Input Multi-class Gaussian Process (NIMGP)[8] (Villacampa-Calvo et al. 2020), which was constructed in such a way to incorporate the uncertainties of the input variables, either given explicitly (say, from the experiment, as it is our case at hand), or to be learned by the model itself. While more details are given in Appendix C3, in short here the idea of such model is to assume that every observation $\mathbf{x}_i$ is a noisy instance of the true value (call it $\tilde{\mathbf{x}}_i$), following a Gaussian distribution.

[7] See Rasmussen & Williams (2006) for the classical textbook about Gaussian Processes.

[8] Here modified to solve a binary classification problem.

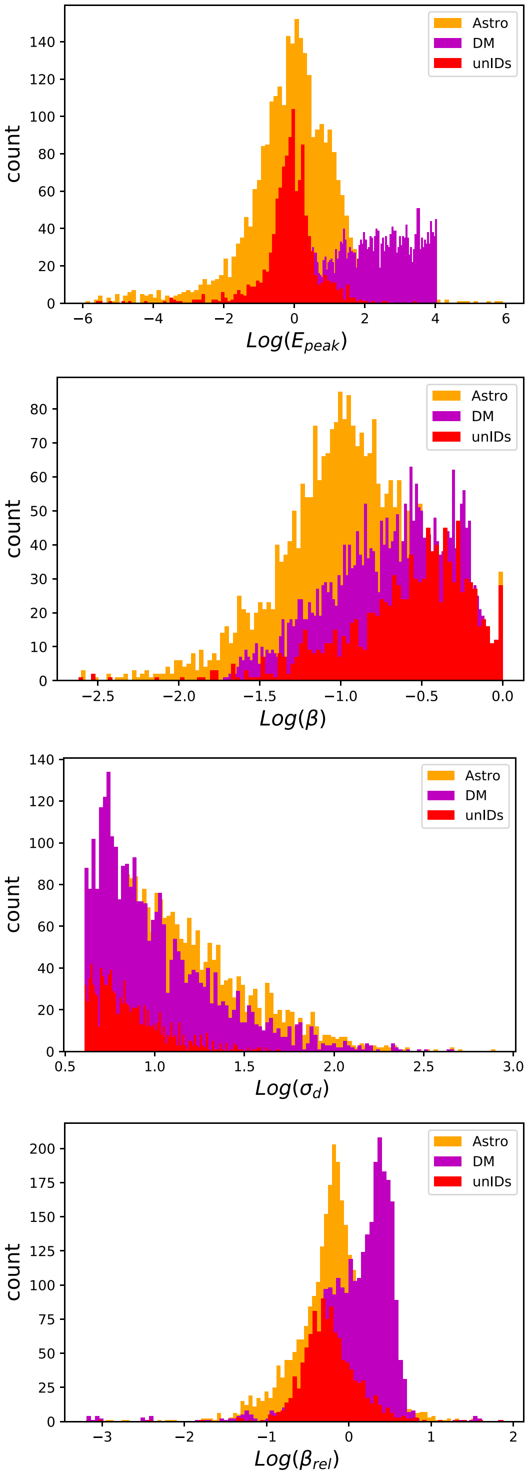

Figure 7. Histograms of the four features of the balanced data[9] adopted. From upper to lower panel: emission energy $E_{\rm peak}$, curvature of the spectra β, detection significance $\sigma_{\rm d}$, relative error on β. In each panel, we show the histograms for the classified astrophysical sources (orange), unIDs (red), and DM data set (magenta).

4.2 Setups

In this work, we consider different setups for the classification task, with increasing level of complexity either from the data set itself as well as from the modelling part. They are described below:

(i) '**2F**'. A data set with only two features: $E_{\rm peak}$ and β. This is the minimum setup, not requiring the construction of additional variables beyond those coming from the fit of the spectrum for both astrophysical sources and DM. This setup does not take into account the uncertainty on β when doing the classification.

(ii) '**3F-A**'. A first strategy for taking into account the uncertainty on β is to follow an heuristic, by which the original data set is artificially augmented, assuming that every observation β_i follows a Gaussian, whose mean is given by the precisely observed value, and the standard deviation is the reported uncertainty on β_i. Then we augment the data set by taking each original point and sampling 60 times from it. Coincides an augmented data set containing three features: $E_{\rm peak}$, $\sigma_{\rm d}$, and $\beta_{\rm sampled}$. More details can be found in the appendix Section B and Fig. B1.

(iii) '**3F-B**'. The second strategy for incorporating the uncertainties in β is inspired in a recent work by Villacampa-Calvo et al. (2020), as commented above and explained in more detail in the appendix. This is arguably the most formal procedure for taking into account the input uncertainties, among all the setups we consider here. The data set here contains the three same features as above, i.e. $E_{\rm peak}$, $\sigma_{\rm d}$, and β. However, now the uncertainties of β are just included in the statistical model. Concretely, this setup will concern exclusively the NIMGP model mentioned above.

(iv) '**4F**'. The last strategy for taking into account the uncertainties of *beta*, is to include them as a separate feature (input variable). This is a priori reasonable, since we have checked that there is only a minor correlation between β and ϵ_β.[10] The data set here contains four features: $E_{\rm peak}$, $\sigma_{\rm d}$, β, and $\beta_{\rm rel}$. Note that, in the case of the DM class, both $\sigma_{\rm d}$ and $\beta_{\rm rel}$ have been constructed out of the unIDs population, as discussed in Section 3.

4.3 Data pre-processing

We pre-process the data as follows:

(i) we apply a cut on $10^{-3} \le E_{\rm peak} \le 10^6$, which is a reliable range of energy due to the Fermi-LAT sensitivity;

(ii) we create the DM data sample in order to have a balanced data set, i.e. the same number of astrophysical (hereafter, astro) and DM data;

(iii) we work in log-space, due to the broad range of values for each feature;

(iv) we standardize data (see e.g. Shanker, Hu & Hung 1996; Bishop 2006), i.e. each feature is transformed to have zero mean and unit variance. For each classification run, the standardization is done with respect to the training data set and testing data set, independently. The unIDs sample has been also standardized.

5 ASTRO-VERSUS-DM CLASSIFICATION RESULTS

In the following, we show the performance of different combinations of the four ML algorithms and setups previously introduced.

First, we consider three models: the LR, NN, and NB models for three of the different setups described in Section 4.2 (the 2F, 3F-A,

[9] balanced data means that the number of data in different classes, here astro and DM, is kept of the same order.

[10] Actually, the Pearson correlation coefficient being equal to 0.4.

Table 1. Performances of the three different ML models LR, NN, NB in the 2F, 3F-A, and 4F setups compared with the GP in the 3F-B setup, as described in 5. See the text for details. We highlight in bold face the results of the configuration giving the best performance.

	OA (per cent)	TN (per cent)	TP (per cent)
LR			
2F	84.9 ± 0.8	85.4 ± 1.5	84.4 ± 1.4
3F-A	83.0 ± 0.1	85.0 ± 0.2	81.0 ±0.2
4F	86.0 ± 0.9	86.7 ± 1.5	85.2 ± 1.3
NN			
2F	86.2 ± 0.8	86.1 ± 3.0	86.4 ± 3.4
3F-A	85.0 ± 0.2	87.9 ± 1.8	82.3 ±1.8
4F	**93.3 ± 0.7**	**94.7 ± 1.7**	**91.8 ± 1.5**
NB			
2F	82.4 ± 1.5	83.9 ± 1.9	80.5 ± 2.5
3F-A	82.5 ± 0.3	83.7 ± 0.4	81.6 ± 0.3
4F	83.5 ± 1.0	86.2 ± 1.2	81.7 ± 1.2
GP			
3F-B	88.1 ± 0.2	89.6 ± 0.3	84.9 ± 0.2

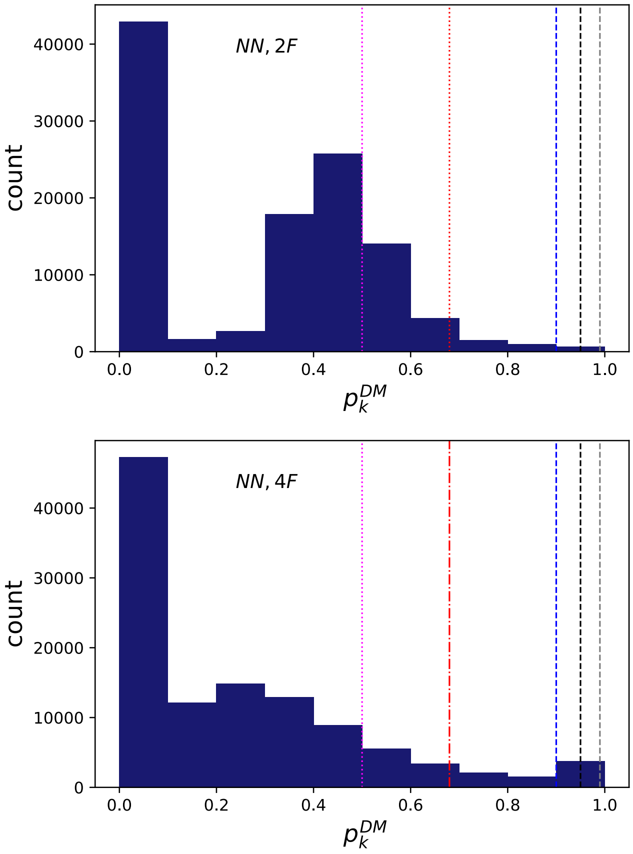

Figure 8. Probability distribution of the full sample of unIDs classified 100 times. The histogram as 100 × $N_{\rm unids}$ entries. The vertical lines correspond to different cut on $p_k^{\rm DM}$, namely 0.50 (magenta-dotted line), 0.68 (red-dotted-dashed line), 0.90 (blue-dashed line), 0.95 (black-dashed line), 0.99 (grey-dashed line). Upper panel: NN classifier in the 2F setup. Lower panel: NN classifier in the 4F setup.

and 4F) and the GP for the 3F-B setup. The results are shown in Table 1, where the columns show the overall classification accuracy (OA), the True Negative (TN) rate, and the True Positive (TP) rate, respectively, while noting that 'negative' here refers to the astro class, while 'positive' refers to the DM class. The False negative (positive) rate can be simply obtained as 100 per cent − TP (TN), being these values normalized over the true. The reported value and quoted uncertainty correspond to the mean and standard deviation of the OA, TN/P rate obtained after 100 splits (see Appendix D for further details). The precision $P = \rm TP/(TP + FP)$ and the False Discovery Rate $\rm FDR = FP/(TP + FP)$ may be also deduced from the table: for the NN-4F we have $P = 0.94 \pm 0.02$ and $\rm FDR = 0.06 \pm 0.02$. We find out that all the classifiers improve their OA, TN, and TP from the 2F to the 4F setup, by including the systematic features. On the other hand, the accuracy decreases for the '3F-A' configuration, for all the three classifiers. The reason for this is simply that, in the augmented data set, the two classes will necessarily overlap more, quantitatively depending on the quoted uncertainty for β. We conclude that among the two strategies considered so far for taking into account $\beta_{\rm rel}$, the 4F setup gives better performance.[11] We get OA = 93.3 per cent ± 0.7 per cent and we can correctly classify 94.7 per cent ± 1.7 per cent of astrophysical sources. Intuitively, we give more importance to the correct classification of already well-known astrophysical sources (TN) than to the one of prospective DM sources (TP), i.e. to a second level, our best classifier will be the one that maximizes not only the OA, but also the TN percentage.

Finally, Table 1 also shows the result of the GP model (specifically, the NIMGP implementation, see Appendix C3) using the 3F-B configuration. Even if this is the more complex model considered (in number of the free parameters of the model to be optimized), we see that its performance concerning classification accuracy is smaller than for the NN model with the 4F setup. This is not surprising: indeed it is common that large Bayesian models may show an overall performance which is not higher than flexible models in a frequentist approach (as the NN). Instead, the advantage of using Bayesian inference comes mainly from its capability of providing estimates of prediction uncertainties, accounting for both statistical (a.k.a aleatoric) and modelling (a.k.a. epistemic) uncertainties. The way to do it is via the predictive distribution $p(y_*|\mathbf{x}_*, \mathcal{D})$, for the class y_* at a new input $\mathbf{x}_*$, given the already observed (training) data $\mathcal{D}$. This is an intrinsically Bayesian quantity, and does not have counterpart with the frequentist implementation of the NN we have adopted in this work.

6 UNIDS CLASSIFICATION

Among our algorithms, we select the one with the best performance – which is the NN with the 4F setup – to classify our unIDs sample and to search for prospective DM-source candidate. None the less – in order to show the improvement in the classification obtained by training the NN with the inclusion of the systematic features – in Fig. 8 we show the classification results obtained from both the 2F and 4F setups. In particular, we show the distribution of probabilities $p_k^{\rm DM}$ of the full sample of unIDs to be classified as DM in each of the $k = 1....100$ classification runs, corresponding to different training/testing split and/or different random seeds. Fig. 8 shows a clear trend in the astro-versus-DM classification of unIDs. On the one hand, many unIDs are classified with probability 30 per cent $\leq p_k^{\rm DM} \leq$ 60 per cent in the 2F setup (upper panel in

[11] We have verified that a further setup with 3F ($E_{\rm peak}$, β, β_{rel}) returns OA= 88.6 per cent ± 0.8 per cent, TN= 87.9 per cent ± 2.8 per cent, TP= 89.2 per cent ± 2.8 per cent, indeed worst than the NN in the 4F setup.

Table 2. Result of classifying the unIDs with the NN model, for the 4F setup considered in this work. The entries represent mean and standard deviation (across the splits) of the number of unIDS (out of 1125 considered in our sample) whose prediction for the probability of being DM is greater than 50 per cent, 68 per cent, 90 per cent, 95 per cent, and 99 per cent, 99 per centrespectively (see also Fig. 9).

Setup	$p_k^{\mathrm{DM}} \geq$ 50 per cent	$p_k^{\mathrm{DM}} \geq$ 68 per cent	$p_k^{\mathrm{DM}} \geq$ 90 per cent	$p_k^{\mathrm{DM}} \geq$ 95 per cent	$p_k^{\mathrm{DM}} \geq$ 99 per cent
4F	162 ± 41	79 ± 35	37 ± 28	27 ± 25	14^{+20}_{-14}

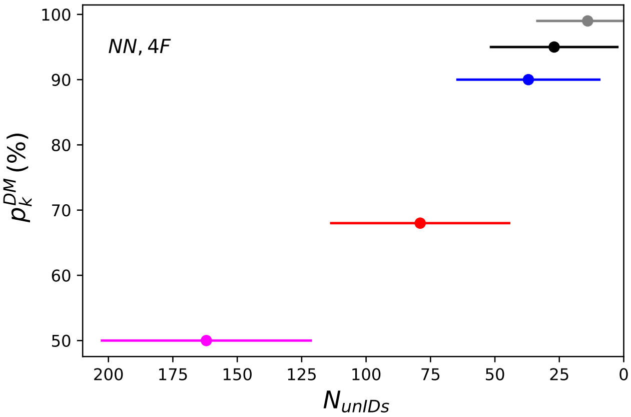

Figure 9. Mean number of unIDs with $p_k^{\mathrm{DM}} > 0.50, 0.68, 0.90, 0.95, 0.99$ in the NN-4F classification. The error bars are calculated as the standard deviation on 100 classifications.

Fig. 8); this peak of probabilities spreads to 0 per cent $\leq p_k^{\mathrm{DM}} \leq$ 60 per cent in the 4F setup, now suggesting that those sources are most probably astrophysical sources. The improvement in such a degeneracy represents a first partial result of this work.

In Table 2 and Fig. 9, we show the mean number of unIDs classified with $p_k^{\mathrm{DM}} \geq$ 50 per cent,68 per cent, 90 per cent, 95 per cent, 99 per cent in each classification, and the standard deviation calculated on $k = 1....100$ classification runs.

Finally, although the number of unIDs classified as DM with $p_k^{\mathrm{DM}} \geq$ 99 per cent is compatible with zero, one may wonder which unIDs of the sample has any $p_k^{\mathrm{DM}} \geq$ 90 per cent. In Fig. 10 we show the counting for each unIDs to be classified with $p_k^{\mathrm{DM}} \geq$ 90 per cent over 100 classifications. We observe that at most 13 out of the 100 classifiers give a probability $p_k^{\mathrm{DM}} \geq$ 90 per cent only for a few unIDs. Due to both such a small counting and the statistical fluctuations, it is indeed impossible to point out a specific best DM candidates among our sample of unIDs, our results being compatible with no DM sources among our unIDs sample.

7 CONCLUSIONS

The main scope of this work was to study the possibility that some of the unidentified gamma-ray point-like sources (unIDs) found at the latest Fermi-LAT catalogue (4FGL) would actually shine due to WIMP DM annihilation.

In order to do that, we first studied the differences between observed astrophysical sources (pulsars, blazars, etc.) and prospective DM candidates in a 2D parameters space which have been shown to offer good discriminatory power in previous studies with astrophysical sources only. Such parameters are E_{peak} and β (Coronado-Blázquez et al. 2019a, b) – defining the so-called β-plot – which among others characterize the energy spectrum of the sources when fitted with a log-parabola shape. We interpreted this problem as a standard machine learning (ML) binary classification task, and considered four ML models for that purpose. In doing this, we found that the above two parameters offered a limited discrimination power, while for the observed astrophysical sources there is much more information available. In particular, two additional parameters, the detection significance σ_d and the uncertainty ϵ_β associated to the β parameter were promising for improving the classification task. In order to use those, we built fictitious values of these two additional parameters for the DM simulated data. We did it by sampling the corresponding distributions of the unIDs, under the main assumption that if WIMPs were actually present, they would show up among the unIDs population. We suggest the inclusion of these synthetic features as an heuristics to easily take into account some of the feature uncertainties in the classification algorithms.

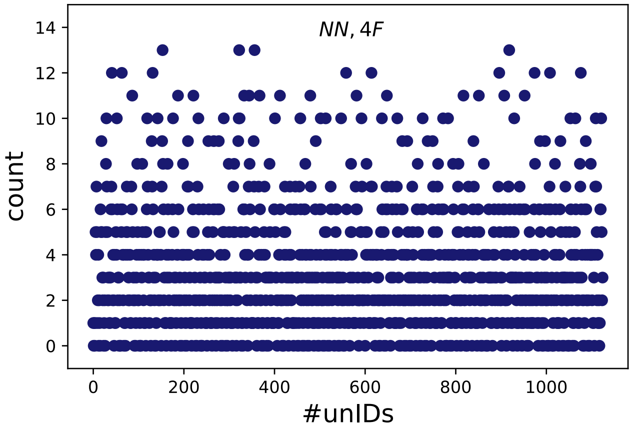

Figure 10. Each point reflects how many times each unIDs is classified with $p_k^{\mathrm{DM}} \geq$ 90 per cent. The best candidates in the run have been classified 13 times over 100 classifications with $p_k^{\mathrm{DM}} \geq$ 90 per cent. None the less, the small counts and the statistical fluctuations with other unIDs do not allow us to claim for a robust DM candidate.

The ML models considered in this work are: Logistic Regression, Neural Network (NN), Naive Bayes, and a particular implementation of a Gaussian Process classifier. The first three models were trained using either two features (E_{peak} and β), three features (including σ_d), or four features (including also ϵ_β). The three-feature setup implemented an augmented data set for taking into account ϵ_β as part of the data itself. The four-feature setup incorporated the information about ϵ_β simply as an extra independent feature instead. The GP model on the other hand, used the same three features as above, while incorporating ϵ_β not as an augmented data set, but as part of the statistical model. Overall, we found that the configuration giving the best performance, in terms of classification accuracy, was the NN with four features (4F), giving a classification accuracy of about 93 per cent ±0.7 per cent.

We selected such setup, NN-4F, as the one for the final task of classifying unIDs as either astrophysical sources or DM sources. We created 100 versions of such a setup, by training the model on 100 data splits (including different train + test partitions as well as random seeds for initializing the weights of the network), thus effectively having 100 predictions for every unID about the probability of being DM. We found that in most cases the predicted probabilities are smaller than 10 per cent, while there is a distribution extending to larger values (cf. Fig. 8 lower panel). Only few unIDs are classified with a larger probability (greater than 90 per cent) to be DM, but only in at most 13 out of the 100 predictions (cf. Fig. 10), while subject to large statistical fluctuations.

We thus conclude that there is no significant evidence for WIMP DM among the unIDs analysed in this work.

As a final word, we would like to remark that – although we found no DM candidates among our sample of LAT unIDs – the proposed methodology appears promising in order to include features uncertainties in classification problems, while with improving the overall performance. In a near future we aim to apply this new methodology. In fact, the proposed methodology is completely model dependent, both on the experimental and theoretical side. On the experimental side, it depends on the characteristics of the instrument, e.g. the energy range, on the theoretical side the WIMP hypothesis could be relaxed searching for e.g. other DM candidates.

ACKNOWLEDGEMENTS

The work of VG and MASC was supported by the grants PROYECTOS DE I+D DE ≪GENERACIÓN DE CONOCIMIENTO PGC2018-095161-B-I00, CENTRO DE EXCELENCIA "SEVERO OCHOA" CEX2020-001007-S, Proyectos I+D+i PID2021-125331NB-I00 all funded by Ministerio de Ciencia e Innovación MCIN/AEI/10.13039/501100011033 and by 'European Regional Development Fund (ERDF) A way of making Europe'. 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 Atracción de Talento de la Comunidad de Madrid under grant no. 2017-T2/TIC-5455, from the Comunidad de Madrid/Universidad Autonoma Madrid 'Proyecto de Jóvenes Investigadores' grant no. SI1/PJI/2019-00294, from Spanish 'Proyectos de I + D de Generacion de Conocimiento' via grant PGC2018-096646-A-I00. BZ also acknowledges the support from Generalitat Valenciana through the plan GenT program (CIDEGENT/2020/055). The work of MASC and JCB was also supported by the Atracción de Talento contracts no. 2016-T1/TIC-1542 and 2020-5A/TIC-19725 granted by the Comunidad de Madrid in Spain.

DATA AVAILABILITY

The *Fermi*-LAT Collaboration follows an open-access policy of data sharing: the data are provided on-line to scientific community at the following link: https://fermi.gsfc.nasa.gov/ssc/data/access/lat/10yr_catalog/.

REFERENCES

Abdollahi S. et al., 2020, ApJS, 247, 33
Ade P. A. R. et al., 2016, A&A, 594, A13
Agarwal S., Davé R., Bassett B. A., 2018, MNRAS, 478, 3410
Aghanim N. et al., 2020, A&A, 641, A6
Ahnen M. L. et al., 2019, MNRAS, 485, 356
Alvarez A., Calore F., Genina A., Read J., Serpico P. D., Zaldivar B., 2020, J. Cosmol. Astropart. Phys., 09, 004
Atwood W. et al., 2013.2012 Fermi Symposium proceedings - eConf C121028, preprint (arXiv:1303.3514)
Bartels R., Edwards T., 2019, Phys. Rev. D, 100, 068301
Bazarov A., Benito M., Hütsi G., Kipper R., Pata J., Põder S., 2022, Astron. Comput., 41, 100667
Belikov A. V., Buckley M. R., Hooper D., 2012, Phys. Rev. D, 86, 043504
Belotsky K., Kirillov A., Khlopov M., 2014, Gravit. Cosmol., 20, 47
Bergström L., 2013, in Walter R., Türler M., eds, Multi-Messenger Astronomy and Dark Matter. Springer, Berlin, Heidelberg, p. 123
Berlin A., Hooper D., 2013, Phys. Rev. D, 89, 095019
Bertone G., Merritt D., 2005, Mod. Phys. Lett. A, 20, 1021
Bertone G., Bozorgnia N., Kim J. S., Liem S., McCabe C., Otten S., Ruiz de Austri R., 2018, J. Cosmol. Astropart. Phys., 03, 026
Bertoni B., Hooper D., Linden T., 2015, J. Cosmol. Astropart. Phys., 12, 035
Bertoni B., Hooper D., Linden T., 2016, J. Cosmol. Astropart. Phys., 5, 049
Bhat A., Malyshev D., 2022, A&A, 660, A87
Bishop C. M., 2006, Pattern Recognition and Machine Learning. Springer, Berlin
Buckley M. R., Hooper D., 2010, Phys. Rev. D, 82, 063501
Calore F., Romeri V. D., Mauro M. D., Donato F., Marinacci F., 2016, Phys. Rev. D, 96, 063009
Calore F., Serpico P. D., Zaldivar B., 2018, J. Cosmol. Astropart. Phys., 10, 029
Caron S., Gómez-Vargas G. A., Hendriks L., Ruiz de Austri R., 2018, J. Cosmol. Astropart. Phys., 05, 058
Cembranos J. A. R., de la Cruz-Dombriz A., Gammaldi V., Lineros R. A., Maroto A. L., 2013, J. High Energy Phys., 09, 077
Charles E. et al., 2016, Phys. Rep., 636, 1
Ciafaloni P., Comelli D., Riotto A., Sala F., Strumia A., Urbano A., 2011, J. Cosmol. Astropart. Phys., 2011, 019
Cirelli M. et al., 2011, J. Cosmol. Astropart. Phys., 03, 051
Conrad J., Reimer O., 2017, Nat. Phys., 13, 224
Coronado-Blázquez J., Sánchez-Conde M. A., Di Mauro M., Aguirre-Santaella A., Ciucă I., Domínguez A., Kawata D., Mirabal N., 2019a, J. Cosmol. Astropart. Phys., 11, 045
Coronado-Blázquez J., Sánchez-Conde M. A., Domínguez A., Aguirre-Santaella A., Mauro M. D., Mirabal N., Nieto D., Charles E., 2019b, J. Cosmol. Astropart. Phys., 2019, 020
Coronado-Blázquez J., Sánchez-Conde M. A., Pérez-Romero J., Aguirre-Santaella A., Fermi-LAT Collaboration, 2022, Phys. Rev. D, 105, 083006
Developers T., 2022, tensorflow. Available at: https://doi.org/10.5281/zenodo.5949169
Feickert M., Nachman B., 2021, GitHub repository of Living Review
Gammaldi V., 2019, Front. Astron. Space Sci., 6, 19
Gammaldi V., Pérez-Romero J., Coronado-Blázquez J., Di Mauro M., Karukes E., Sánchez-Conde M. A., Salucci. P., 2021, PoS, ICRC2021, 509
Germani S., Tosti G., Lubrano P., Cutini S., Mereu I., Berretta A., 2021, MNRAS, 505, 5853
Holwerda B. W. et al., 2022, MNRAS, 513, 1972
Hooper D., Witte S. J., 2017, J. Cosmol. Astropart. Phys., 4, 018
Hui C. Y. et al., 2020, MNRAS, 495, 1093
Kovačević M., Chiaro G., Cutini S., Tosti G., 2019, MNRAS, 490, 4770
Mirabal N., 2013, MNRAS, 436, 2461
Mirabal N., Charles E., Ferrara E. C., Gonthier P. L., Harding A. K., Sánchez-Conde M. A., Thompson D. J., 2016, Am. Astron. Soc., 825, 69
Morice-Atkinson X., Hoyle B., Bacon D., 2018, MNRAS, 481, 4194
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Peebles P. J. E., Schramm D. N., Turner E. L., Kron R. G., 1991, Nature, 352, 769
Rasmussen C. E., Williams C. K. I., 2006, Gaussian Processes for Machine Learning. MIT Press, Cambridge
Schoonenberg D., Gaskins J., Bertone G., Diemand J., 2016, J. Cosmol. Astropart. Phys., 5, 028
Shanker M., Hu M., Hung M., 1996, Omega, 24, 385
Spencer S., Armstrong T., Watson J., Mangano S., Renier Y., Cotter G., 2021, Astropart. Phys., 129, 102579
The Fermi-LAT Collaboration, 2012, ApJ, 750, 3
The Fermi-LAT Collaboration, 2015, ApJS, 218, 23
The Fermi-LAT Collaboration, 2016, ApJS, 222, 5
The Fermi-LAT Collaboration, 2017, ApJS, 232, 18
Ullmo M., Decelle A., Aghanim N., 2021, A&A, 651, A46
Villacampa-Calvo C., Zaldivar B., Garrido-Merchán E. C., Hernández-Lobato D., 2020, J. Mach. Learn. Res., 22, 1
Visinelli L., 2018, Symmetry, 10, 546
Zechlin H. S., Fernandes M. V., Elsaesser D., Horns D., 2011, A&A, submitted
Zechlin H.-S., Horns D., 2012, JCAP, 11, 050

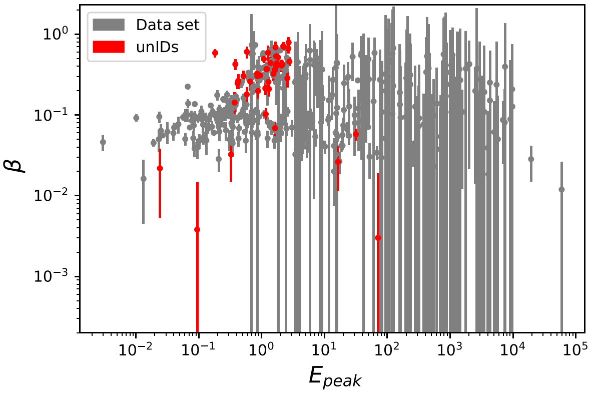

Figure A1. Same as Fig. 1, by including the systematic features created for the DM data set. A cut on $\sigma \geq 20$ is applied for abetter clarity of the plot.

APPENDIX A: DM–β PLOT WITH SYSTEMATIC FEATURE ON β

In Fig. A1 we show an example of the β-plot with the astrophysical and DM data set including the systematic uncertainty on β. A cut on $\sigma \geq 20$ is applied for the clearness of the plot. The uncertainty on the data set is only partially correlated to the detection significance, accordingly with the observational sample of unIDs sources.

APPENDIX B: SAMPLED GUASSIAN DISTRIBUTION OF β UNCERTAINTY

In this Appendix we describe a different methodology to include the uncertainty on β in the classification algorithm. Instead of incorporating the uncertainty of β as an extra feature, another option is to include it implicitly in the data by creating an augmented data set $\beta_{\rm sampled}$. The strategy here is to augment the data set for each observation $i = 0,...N$, we assume that the variable β follows a truncated Gaussian distribution, whose mean is precisely the observed value β_i, and the standard deviation is precisely the observed value ϵ_{β_i}, but truncated such that $0 < \beta \leq 1$. We then sample $M = 60$ points from such truncated Gaussian, such that we obtain an augmented data set of size $M \cdot N$. This is an heuristic methodology for taking into account the uncertainties in the input variables, by using only three features (3F-A) $E_{\rm peak}$, $\beta_{\rm sampled}$, σ. We show both the 'DM-β' plot and the final histograms for each feature in Fig. B1. None the less, we found some issues with this methodology. First, the augmented number of data makes the classification slower. Secondly, the augmented data set imply a substantially larger overlap between the two classes of points (both in the $DM - \beta$ plot and histograms of each features, see Fig. B1), with the consequent loss in discrimination power. Thirdly, this method requires the use of only three features in the learning process. Indeed, the same algorithm applied to the unIDs classification, implies the use of only three features, with two options: (1) neglecting the $\beta_{\rm rel}$ for the unIDs and using only the mean value of β as feature, so preventing us from using the available information contained in the unID's $\beta_{\rm rel}$; (2) including $\beta_{\rm rel}$ by augmenting the unIDs sample with the same Gaussian sample methodology, which at the end will bring to an overlapping of different unIDs that would be very hard to reconstruct. For all these reasons, we do not use this methodology for the unIDs classification.

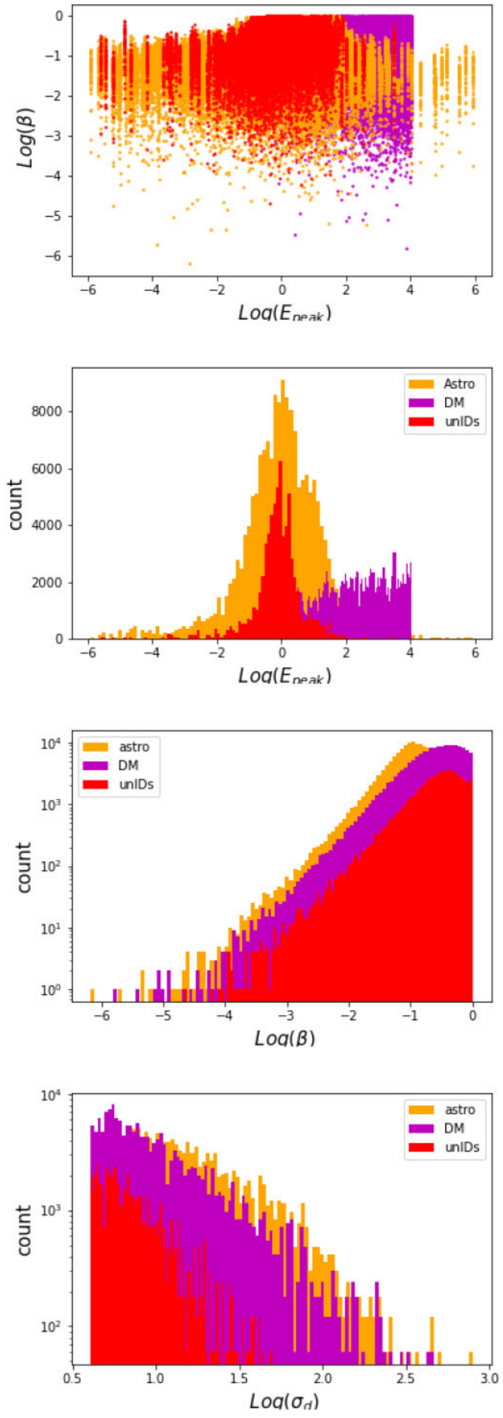

Figure B1. DM-β plot (upper panel) and histograms of the three features of the augmented data set. From the second upper to lower panel: characteristic emission energy $E_{\rm peak}$, curvature of the spectra $\beta_{\rm sampled}$, detection significance $\sigma_{\rm d}$. In this set-up, the relative error is included as the standard deviation of a truncated Gaussian around the mean value β, i.e. we only have three features (3F-A, as explained in the B).

APPENDIX C: TECHNICAL DETAILS OF THE ML MODELS

For the sake of reproducibility, we specify in this section the implementation details of some of the ML models considered in this work.

C1 Logistic regression

We use the implementation of LR as given in the python library `scikit-learn` (Pedregosa et al. 2011), indeed class:

sklearn.linear_model.LogisticRegression (penalty='l2', ∗, dual=False, tol=0.0001, C = 1.0, fit_ intercept = True, intercept_ scaling=1, class_ weight = None, random_ state=None, solver = 'lbfgs', max_ iter = 100, multi_ class='auto', verbose = 0, warm_ start = False, n_ jobs = None, l1_ ratio = None)

C2 Neural network

We use the implementation of NN as given in the python library `scikit-learn` (Pedregosa et al. 2011). Specifically, we use the `MLPClassifier` model, with the following setup:

sklearn.neural_network.MLPClassifier(hidden_layer_sizes=(21,), activation='relu', ∗, solver='adam', alpha = 0.0,batch_size = 120, learning_rate = 'constant',learning_rate_init = 0.0015,power_t = 0.5, max_iter=1000,shuffle = True,random_state=None,tol=0.0001, verbose = False,warm_start=False,momentum = 0.9, nesterovs_momentum = True,early_stopping = False, validation_fraction = 0.1,beta_1 = 0.9, beta_2=0.999,epsilon = 1e-08, n_iter_no_change = 10,max_fun = 15000).

In Fig. C1 we show the performance of the NN for one and two layers with different number of neurons. We choose to use a 1-layer configuration with 21 neurons, due to the combination of good accuracy (upper panel) and low fitting time (lower panel).

C3 Gaussian process

As for the previous models, in this case we also work in log-space of the attributes. However, care must be taken in this case, since then the uncertainties of $\log\beta$ should be computed properly. This is a standard procedure, which however is depicted next:

The idea is to compute the uncertainties of a new variable $\ell \equiv \log\beta$, where $\beta \sim \mathcal{N}(\beta|\mu_\beta, \sigma_\beta)$, while the given uncertainties in β are assumed to be precisely its standard deviation, σ_β. In order to do this, we take as the ℓ's uncertainties to be the square root of its variance, computed from its own pdf:

$$p(\ell|\mu_\beta, \sigma_\beta) = e^{\ell}\mathcal{N}(e^{\ell}|\mu_\beta, \sigma_\beta)\,. \tag{C1}$$

We have adapted to our case one of the variants of the Gaussian Process (GP) models presented in Villacampa-Calvo et al. (2020): the NIMGP$_{\rm NN}$ model, which is roughly described next, while for further details we refer the reader to Villacampa-Calvo et al. (2020).

The NIMGP$_{\rm NN}$ model assumes a likelihood for the label y_i of the i-th point as follows:

$$\begin{aligned} p(y_i|\mathbf{f}_i) = {} & (1-\epsilon)\prod_{c\neq y_i}\Theta\left(f^{y_i}(\mathbf{x}_i) - f^{c}(\mathbf{x}_i)\right) \\ & + \frac{\epsilon}{C-1}\left[1 - \prod_{c\neq y_i}\Theta\left(f^{y_i}(\mathbf{x}_i) - f^{c}(\mathbf{x}_i)\right)\right]\,, \end{aligned} \tag{C2}$$

where $\mathbf{f}_i \equiv \{f^c(\mathbf{x}_i)\}$, $c = 1,.., C$, are the corresponding values of the GP for all C classes, evaluated at the latent inputs $\mathbf{x}_i$ (see below). We account for the possibility of having mislabelled classes by having a small probability $\epsilon = 0.001$ for mislabelling. Finally, $\Theta(\cdot)$ is the Heaviside step function. Note that this is not the typical likelihood considered in popular classification tasks, which correspond to the cross-entropy loss function. However, this is a common choice in the GP classification context, with the added value of accounting for mislabelling errors, something which is typically not taking into account in the cross-entropy setup.

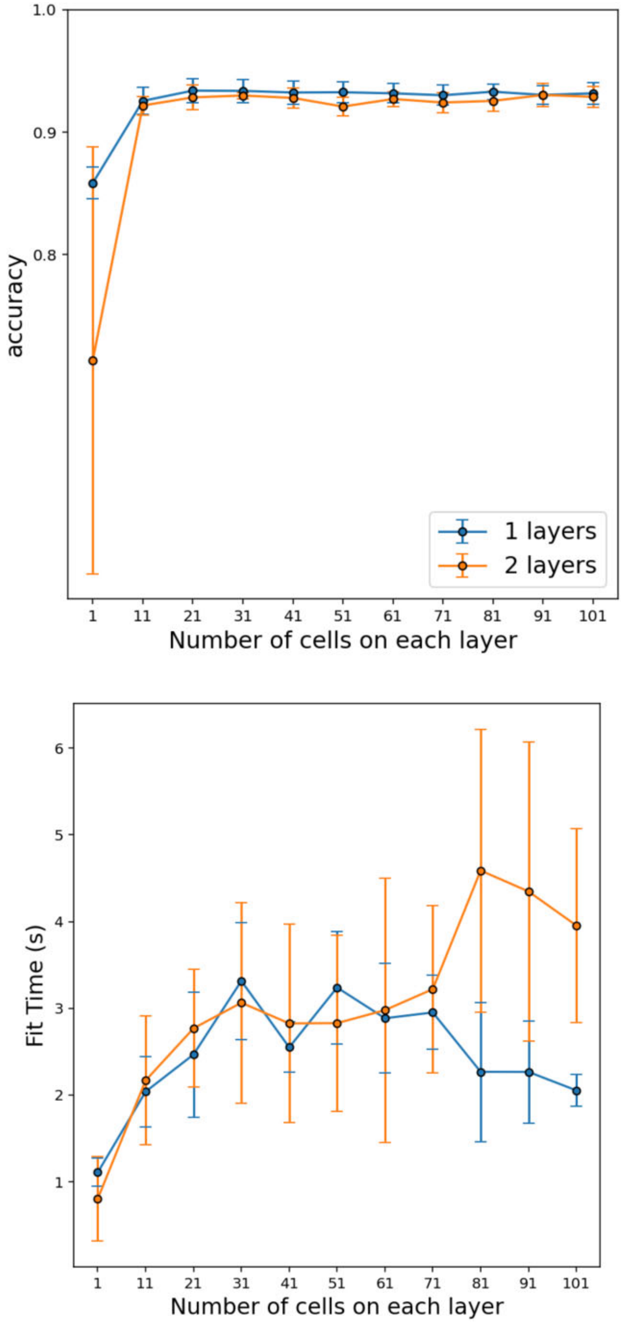

Figure C1. Performance comparing 1-hidden-layer and 2-hidden-layer neural network for the 4F data set. Upper panel: Overall accuracy as a function of the number of neurons in each layer. Lower panel: Time of the learning procedure as a function of the number of neurons in each layer. In the following, we use 1 layer and 21 neurons.

In NIMGP$_{\rm NN}$, it is assumed that the observed input $\tilde{\mathbf{x}}_i$ is a noisy realization of the true (but latent) input $\mathbf{x}_i$, according to the distribution:

$$p(\tilde{\mathbf{x}}_i|\mathbf{x}_i) = \mathcal{N}(\tilde{\mathbf{x}}_i|\mathbf{x}_i, \sigma_i)\,. \tag{C3}$$

As it is well-known, the posterior distribution $p(\mathbf{f}|\mathbf{y})$ of the GP values $\mathbf{f}$ has a computational cost of $\mathcal{O}(N^3)$, where N is the number of points, so this setup is not scalable to very large data sets. For that reason the NIMGP$_{\rm NN}$ adopts a 'Sparse GP' configuration where inference is done only in a subset $\mathbf{u}$ of GP values at some given inputs, called in the literature 'inducing points', and there are $M < N$ of them. Consequently, the latent variables of the model can be grouped in three matrices: (i) $\mathbf{F}$, the $N \times C$ matrix of process values $\mathbf{f}_i$ at the datapoints, (ii) $\mathbf{U}$, the $M \times C$ matrix of process values $\mathbf{u}_j$ at the inducing points, and (iii) $\mathbf{X}$, the $N \times D$ matrix of latent inputs $\mathbf{x}_i$.

The posterior distribution for the above latent variables is intractable, as typical in Bayesian inference, and NIMGP$_{\rm NN}$ approximates it by Variational Inference, where the approximate distribution $q(\mathbf{x}_i)$ is taken as:

$$q(\mathbf{x}_i) = \mathcal{N}\left(\mathbf{x}_i|\boldsymbol{\mu}_\theta(\tilde{\mathbf{x}}_i, y_i), \mathbf{V}_\theta(\tilde{\mathbf{x}}_i, y_i)\right),$$

where both $\boldsymbol{\mu}_\theta(\tilde{\mathbf{x}}_i, y_i)$ and $\mathbf{V}_\theta(\tilde{\mathbf{x}}_i, y_i)$ are obtained as the output of a neural network with parameters θ.

The final scope in Bayesian inference is to compute the predictive distribution $p(y_*|\mathbf{x}_*, \mathcal{D})$, for the class y_* at a new input $\mathbf{x}_*$, given the already observed (training) data $\mathcal{D}$, which in the case of GP binary classifier is given by:

$$p(y_*|\mathbf{x}_*, \mathcal{D}) = \int df_* p(y_*|f_*) p(f_*|\mathbf{x}_*, \mathcal{D}) \,, \tag{C4}$$

where f_* is the process value at the input $\mathbf{x}_*$, while

$$p(f_*|\mathbf{x}_*, \mathcal{D}) = \int d\mathbf{f}\; p(f_*|\mathbf{x}_*, \mathbf{f}) p(\mathbf{f}|\mathcal{D}) \,, \tag{C5}$$

being $p(\mathbf{f}|\mathcal{D})$ the posterior distribution of the process values at all the training points.

As a final note, the NIMGP$_{\rm NN}$ model is written in `python 2` with `tensorflow 1` (see github repository at Villacampa-Calvo et al. 2020).

APPENDIX D: NUMBER OF FOLDS

For the LR and NN, we use the Repeated Stratified K-Fold cross validator, class `RepeatedStratifiedKFold(n_splits = 5, n_repeats = 20)` defined in `scikit-learn` (Pedregosa et al. 2011). By splitting the data in fivefolds, we take 80 per cent of data for the training set and 20 per cent of data for the testing set. This choice allows us to preserve the independence of the five testing sets, i.e. without any repetition of same data. In order to preserve this characteristic, the ratio of testing/training data decreases by increasing the number of folds, while the accuracy of the classification decreases (Fig. D1). Such a split, is repeated 20 times with different random seeds. In this way we have a total of 100 classifications, which allow us to get the results with a reliable statistical uncertainty.

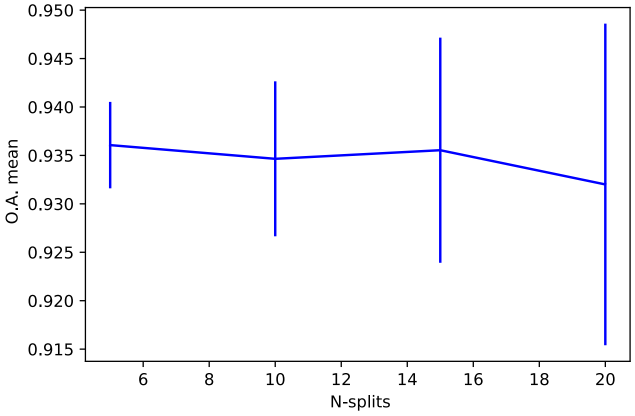

Figure D1. Performance comparing the OA of the NN for different *N*-folds, for the 4F data set. See Section 5 for details.

The uncertainty related with the number of folds is also shown in Fig. D2, where we show the ROC (Receiver Operating Characteristics) and AUC (Area Under the Curve) of the NN for different fivefolds, for the 4F data set.

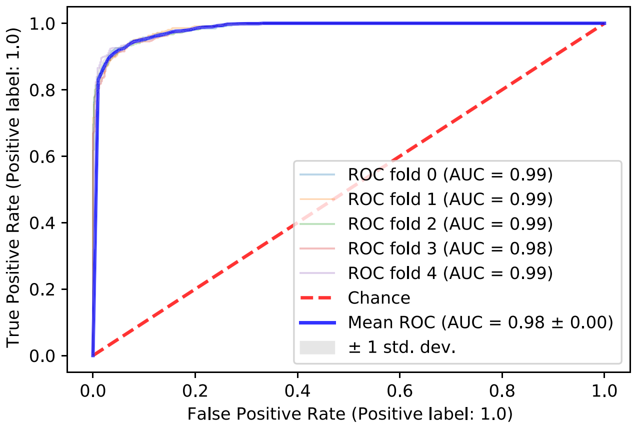

Figure D2. Performance comparing the ROC (Receiver Operating Characteristics) and AUC (Area Under the Curve) of the NN for different fivefolds, for the 4F data set. A model with perfect performance will have AUC = 1, which means it has a good measure of separability. The red dashed line indicates the worst ROC situation, where the AUC = 0.5, the model has no discrimination capacity to distinguish between positive class and negative class.

APPENDIX E: CLASSIFICATION WITH 2-FEATURES

In Figs E1 and E2, we show the same analysis of the algorithm performances for the classification with two-features, indeed without taking into account the systematic features. The performance of the algorithm improves from 2F to 4F both in terms of overall accuracy and ROC/AUC.

In Table E1 and Fig. 8 of this Appendix we show the results of the unIDs classification for the NN in the 2F setup. In Fig. 7 in the main text, we show the probability distribution for the full set of unIDs and 100 classification runs. The improvement in the overall classification behaviour of the 4F setup is visible by comparing this figure with the same figure for the 4F setup, presented in the main text. In the upper panel in Fig. E3, we show the mean number of unIDs classified with $p_k^{\rm DM} > 0.5, 0.68, 0.90, 0.95, 0.99$ and their standard deviation. Finally, in the last panel we showed the count for each unIDs to be classified with $p_k^{\rm DM} \geq 0.90$. The best candidates are classified five times over 100 for the 2F setup (and 13 times over 100 in the 4F setup presented in the main text). Although the number of unIDs with $p_k^{\rm DM} \geq 50$ per cent decreases a 27 per cent from the 2F to the 4F setups, the number of unIDs with $p_k^{\rm DM} \geq 68$ per cent increases by 50 per cent. Yet the number of unIDs with $p_k^{\rm DM} \geq 99$ per cent is compatible with zero for both the 2F and 4F setups. As in the 4F setup, the statistical fluctuations prevent us from claiming any robust DM candidate among the unIDs of the 4FGL Fermi-LAT catalogue.

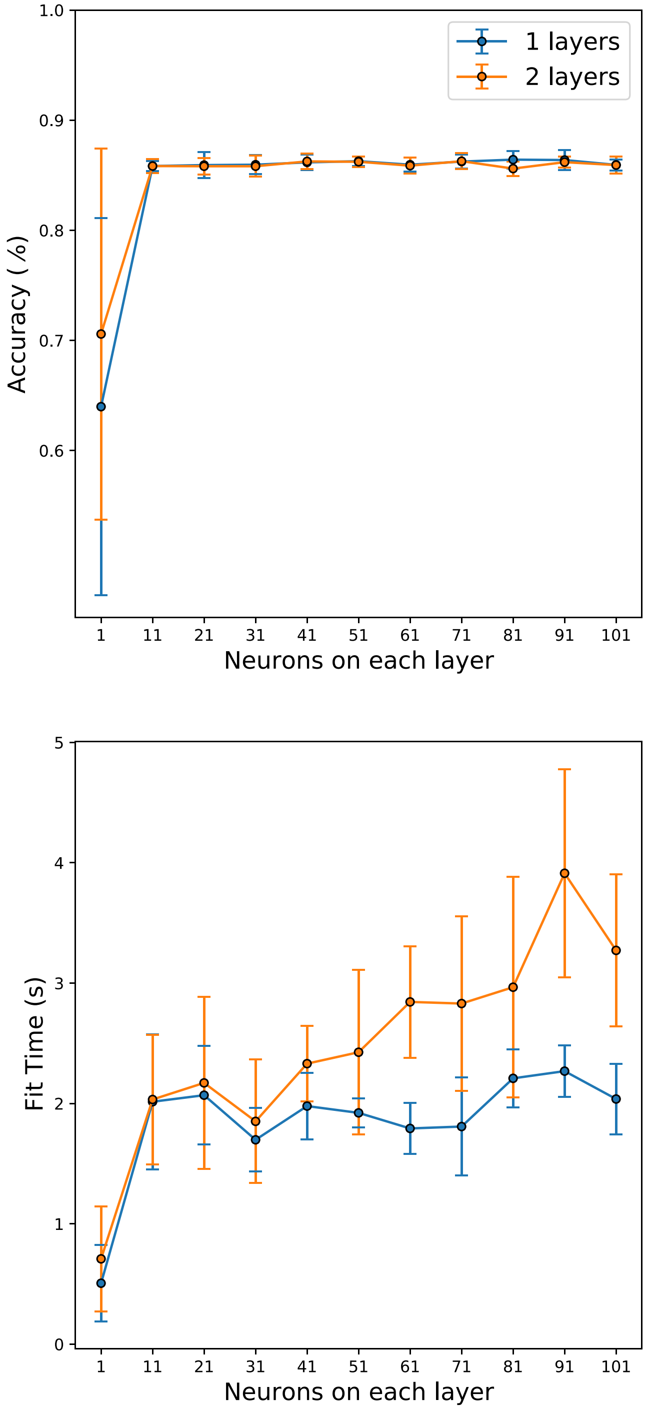

Figure E1. Same as Fig. C1 for the 2F classification.

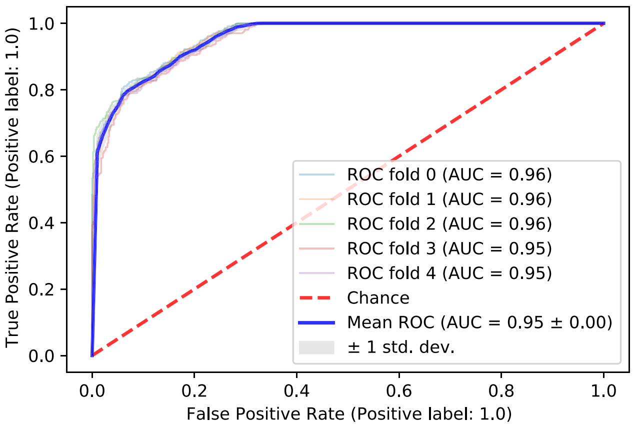

Figure E2. Same as in Fig. D2 for the classification without systematic features.

Table E1. Same as Table 2 for the NN in the 2F setup. See also the upper panel in Fig. E3.

Setup	$p_k^{\mathrm{DM}} \geq$ 50 per cent	$p_k^{\mathrm{DM}} \geq$ 68 per cent	$p_k^{\mathrm{DM}} \geq$ 90 per cent	$p_k^{\mathrm{DM}} \geq$ 95 per cent	$p_k^{\mathrm{DM}} \geq$ 99 per cent
2F	215 ± 45	35 ± 21	6^{+10}_{-6}	3^{+7}_{-3}	1^{+5}_{-1}

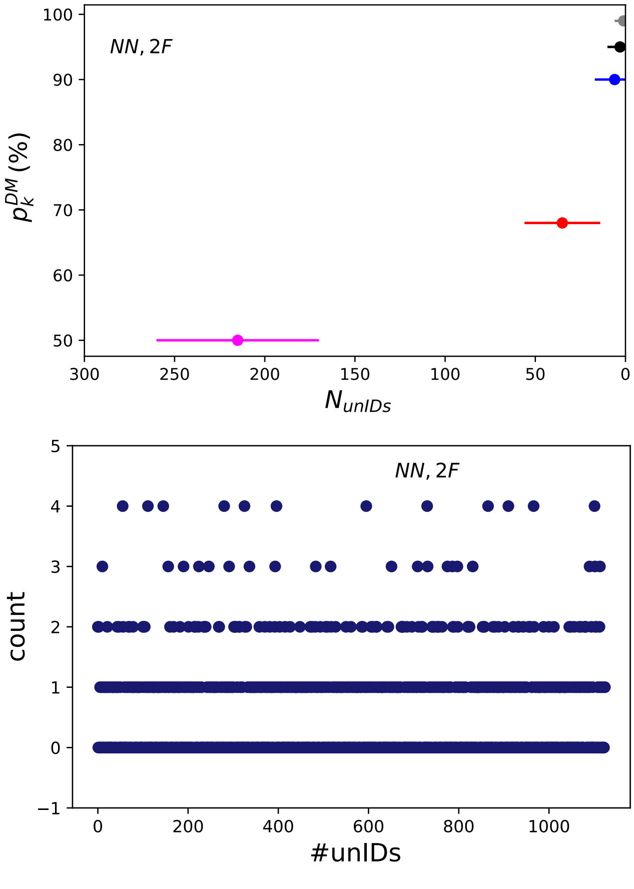

Figure E3. Same as Figs 9, 10 for the NN-2F classification.

This paper has been typeset from a TEX/LATEX file prepared by the author.