Astronomy & Astrophysics manuscript no. 0head_ML_3FGL
December 30, 2021

Machine learning methods for constructing probabilistic *Fermi*-LAT catalogs

A. Bhat [⋆1] and D. Malyshev [⋆⋆1]

Erlangen Centre for Astroparticle Physics, Erwin-Rommel-Str. 1, Erlangen, Germany

December 30, 2021

ABSTRACT

Context. Classification of sources is one of the most important tasks in astronomy. Sources detected in one wavelength band, for example using gamma rays, may have several possible associations in other wavebands or there may be no plausible association candidates.

Aims. In this work, we aim to determine probabilistic classification of unassociated sources in the third and the fourth data release 2 *Fermi* Large Area Telescope (LAT) point source catalogs (3FGL and 4FGL-DR2) into two classes (pulsars and active galactic nuclei (AGNs)) or three classes (pulsars, AGNs, and other sources).

Methods. We use several machine learning (ML) methods to determine a probabilistic classification of *Fermi*-LAT sources. We evaluate the dependence of results on meta-parameters of the ML methods, such as the maximal depth of the trees in tree-based classification methods and the number of neurons in neural networks.

Results. We determine a probabilistic classification of both associated and unassociated sources in 3FGL and 4FGL-DR2 catalogs. We cross-check the accuracy by comparing the predicted classes of unassociated sources in 3FGL with their associations in 4FGL-DR2, for cases where such associations exist. We find that in the 2-class case it is important to correct for the presence of other sources among the unassociated ones in order to realistically estimate the number of pulsars and AGNs. We find that the 3-class classification, in spite of different types of sources in the OTHER class, has similar performance as the 2-class classification in terms of reliability diagrams and, at the same time, it does not require adjustment due to presence of the OTHER sources among the unassociated sources. We show an example of the use of the probabilistic catalogs for population studies, which include associated and unassociated sources.

Key words. Methods: statistical – Catalogs – Gamma rays: general

1. Introduction

Multi-wavelength association of astronomical sources is important for understanding their nature. Unfortunately, in many cases a firm association of sources at different wavelengths is not possible. For example, about one third of the gamma-ray sources in *Fermi* Large Area Telescope (LAT) catalogs are unassociated (Abdo et al. 2010a; Nolan et al. 2012; Acero et al. 2015; Abdollahi et al. 2020). It is at least useful to know the classes to which the unassociated sources belong to or, as is more typical, the probabilities for the sources to belong to various classes. In this paper we use several machine learning (ML) algorithms to find a probabilistic classification of sources in the third *Fermi*-LAT catalog (3FGL, Acero et al. 2015) and the fourth data release two catalog (4FGL-DR2, Abdollahi et al. 2020; Ballet et al. 2020). We used the versions gll_psc_v16.fit for 3FGL and gll_psc_v27.fit for 4FGL-DR2.

We will refer to the catalogs, where the classification of the sources is given in terms of probabilities as probabilistic catalogs. In general, the classes may include the possibility that a source is not a real source but a fluctuation of the background (Caron et al. 2021) or that a source is an overlay of two sources. Probabilistic catalogs have previously been introduced for optical sources (e.g., Hogg & Lang 2010; Brewer et al. 2013) and for gamma-ray sources (Daylan et al. 2017). Bayesian association probabilities were also included in the 4FGL (Abdollahi et al. 2020) and 4FGL-DR2 (Ballet et al. 2020) catalogs for faint sources. Probabilistic classification of unassociated *Fermi*-LAT sources was performed, for example, by Ackermann et al. (2012); Saz Parkinson et al. (2016); Mirabal et al. (2016); Lefaucheur & Pita (2017); Luo et al. (2020); Finke et al. (2020); Zhu et al. (2021),

⋆ e-mail: aakash.bhat@fau.de
⋆⋆ on leave of absence from NRC "Kurchatov Institute" - ITEP, B. Cheremushkinskaya st. 25, Moscow, Russia 117218, e-mail: dmitry.malyshev@fau.de

or in the application for sub-classification of blazars by Hassan et al. (2013); Doert & Errando (2014); Chiaro et al. (2016); Salvetti et al. (2017); Kovačević et al. (2019, 2020) and in subclassification of pulsars by Lee et al. (2012); Saz Parkinson et al. (2016). In this work, we consider the classification of gamma-ray sources into two classes (AGNs and pulsars) as well as into three classes (AGNs, pulsars, and other associated sources). We revisit probabilistic classification of 3FGL sources and compare the results of the classification of unassociated sources with their respective associations in 4FGL-DR2. We also determine a probabilistic classification of the 4FGL-DR2 sources.

Catalogs of gamma-ray point sources are typically designed to have low false detection rates. Nevertheless, 469 sources out of the 3033 in the 3FGL catalog (Acero et al. 2015) have no counterparts in the 4FGL catalog (Abdollahi et al. 2020). This is much larger than the expected false detection rate in 3FGL arising from statistical fluctuations. For the majority of sources in the 3FGL catalog without counterparts in the 4FGL catalog the problem is not the false detection, but rather the association. For example, some sources can be detected due to deficiencies in the Galactic diffuse emission model. In this case, the statistical significance of the detection is high, but the association is wrong: the sources should be classified as a part of the Galactic diffuse emission rather than point-like sources. Another reason could be that two (or more) point-like sources in 3FGL are associated to a single extended source in 4FGL, or a single source is resolved into two sources. Again, this is a problem of classification (or association) rather than false detection.

Another reason for the absence of a previously detected source in a new catalog is variability. In particular, flat spectrum radio quasars (FSRQs) are highly variable AGNs. If a source was active during the observation time of 3FGL but inactive afterwards, then its significance in the 4FGL can be below the detection threshold. This problem is connected to a selection of a hard detection threshold of $TS = 25$ for 3FGL and 4FGL catalogs. Selection of a lower detection threshold could help to keep the variable sources inside the catalog, but it will not solve the problem, since the variable sources near the lower threshold can also disappear in the new catalog. Moreover, a lower threshold would lead to more false detections due to fluctuations of the background. Thus, on the one hand, a lower threshold can be useful in studies where a more complete list of sources is desirable, while the higher false detection rate is admissible. On the other hand, a lower threshold can be problematic for studies where a clean sample is necessary. The problem of the detection threshold selection can be ameliorated with the development of a probabilistic catalog. In this catalog, each point-like object detected above a certain relatively low confidence level is probabilistically classified into classes, which include the statistical fluctuation class. At low confidence, the probability for a source to come from a background fluctuation is high. This probability decreases as the significance of sources increases. Apart from the statistical fluctuation class, classes can include various types of Galactic and extragalactic sources, diffuse emission deficiencies, extended sources, etc. Any user of such a catalog has the freedom to choose the probability threshold for the class that he or she is interested in. In this paper we make a first step in this direction by providing a probabilistic classification of *Fermi*-LAT sources into two or three classes. We also show how the probabilistic catalogs can be used for population studies of sources, e.g., as a function of their flux or position on the sky, where one includes not only associated sources but also unassociated ones according to their class probabilities.

The paper is organized as follows. In Section 2 we discuss general questions about construction of the probabilistic catalogs and the choices of the ML methods. In Section 3 we construct the classification algorithms using the associated sources in the 3FGL catalog for training. We consider several aspects: 1) feature selection, 2) training of the algorithms and selection of meta-parameters, 3) oversampling of the datasets in order to have equal number of pulsars and AGNs in training (there are many more AGNs observed than pulsars).

In Section 4 we apply the classification algorithms determined in Section 3 for the classification of 3FGL and 4FGL-DR2 sources. We compare our predictions for the unassociated sources in 3FGL with the respective associations in 4FGL-DR2. In Section 5 we classify sources in the 3FGL and 4FGL-DR2 catalogs into three classes (AGNs, pulsars, and other sources). In Section 6 we show applications of the probabilistic catalogs for predicting the number of pulsars, AGNs, and other sources among the unassociated sources and in construction of the source counts as a function of their flux, $N(S)$, and as a function of Galactic latitude and longitude, $N(b)$ and $N(\ell)$. We compare the $N(S)$, $N(b)$, and $N(\ell)$ distributions for associated and unassociated sources in the 3FGL and 4FGL-DR2 catalogs. In Section 7 we present our conclusions.

In Appendix A we perform further studies of meta-parameters of some of the ML algorithms, in Appendix B we compare the oversampling method used in the paper with the SMOTE oversampling, while in Appendix D we discuss the reliability diagrams.

2. Choice of methods

2.1. General methodology

The first choices that must be made to construct probabilistic catalogs are the choices of the input data and the machine learning methods to be used. For the input data we take associated point sources (PS) in the 3FGL or 4FGL-DR2 catalogs, which we then split into training and testing subsets. We consider four machine learning algorithms: random forests (RF, Ho 1998; Breiman 2001), boosted decision trees (BDT, Friedman 2001a), logistic regression (LR, Cox 1958), and neural networks (NN, Hopfield 1982). Although the performance of algorithms on testing data is slightly different, we report the classification probabilities for all four algorithms. The difference among the predictions serves as a measure of modeling uncertainty related to the choice of the classification algorithm.

2.2. Discussion of the classification algorithms

One of the most simple and transparent algorithms for classification is decision trees. In this algorithm, at each step the sample is split into two subsets using one of the input features. The choice of the feature and the separating value are determined by minimizing an objective function, such as misclassification error, Gini index, or cross-entropy. This method is very intuitive, since at each step the results can be described in words. For example, at the first step, the sources can be split into mostly Galactic and extragalactic sources by a cut on the Galactic latitude. At the next step, the high latitude sources can be further sub-split into millisecond pulsars and other sources, by a cut on the spectral index around 1 GeV (pulsars have a hard spectrum below a few GeV), etc. One of the main problems with decision trees is either overfitting or bias: if a tree is too deep, then it will pick up particular cases of the training sample resulting in overfitting, while if the trees are too shallow they will not be able to describe the data well, thereby leading to a bias. As a result, one needs to be very careful when selecting the depth of the tree. This problem can be avoided if a random subset of features is used to find a division at each node. This is the basis of the RF algorithm, where the final classification is given by an average of several trees with random subsets of features used at each node. Another problem with the simple trees algorithm is that it can miss the classification of some subsets of data. This is rectified in the BDT algorithm, where the final classification is given by a collection of trees, where each new tree is created by increasing the weights of misclassified samples of the previous step. Finally, simple trees predict classes for the data samples, while we would like to have probabilities for these classes (also known as soft classification). RF and BDT algorithms, by virtue of averaging, provide probabilities. As a result, we will use RF and BDT algorithms rather than simple decision trees in this paper.

Tree-based algorithms, even after averaging in RF and BDT methods, have sharp edges among domains with different probabilities. In LR algorithm, the probabilities of classes are by construction smooth functions of input features. In particular, for two-class classification the probability of class 1, given the set of features x, is modeled by the sigmoid (logit) function

$$p_1(x) = \frac{e^{m(x)}}{1 + e^{m(x)}}. \quad (1)$$

The probability of class 0 is then modeled as $p_0(x) = 1 - p_1(x)$. Therefore, if $m(x)$ is a linear function of features, then the boundary between the domains, defined, e.g., as $p_1(x) = 0.5$, will also be linear at $m(x) = 0$. More complicated boundaries can be modeled by taking non-linear functions $m(x)$. Unknown parameters of the function $m(x)$ are determined by maximizing the log likelihood of the model given the known classes of the data in the training sample. A useful feature of the LR method is that it, by construction, provides probabilities of classes with smooth transitions among domains of different classes. A limitation is that the form of the probability function is fixed to the sigmoid function in Eq. (1).

We notice that if $m(x)$ is a linear function of features x, then the LR model is obtained by an application of sigmoid function to a linear combination of input features. This is in fact a single layer perceptron, or a NN, with several input nodes (each node corresponding to a feature) and one output node, which corresponds to $p_0(x)$, but without any hidden layers. The output value is obtained by a non-linear transformation (sigmoid) of a linear combination of features. A neural network with several hidden layers is obtained by a sequence of non-linear transformations of linear combinations of features. In particular, the values in the first hidden layer are obtained by a non-linear transformation of linear combinations of input features. Then the values in the second hidden layer are obtained by a non-linear transformation of linear combinations of values in the first hidden layer and so on till the required number of hidden layers is reached. In the context of neural networks, the non-linear transformations are also called activation functions. If the activation function for the output layer is sigmoid, then the output values can be interpreted as probabilities.

3. Construction of probabilistic catalogs

One of the first problems one has to deal with for the 3FGL and 4FGL-DR2 catalogs, is that some of the sources in the catalogs have missing or unphysical values (e.g., infinity). In order to avoid a bias in predictions, we include sources with missing or unphysical values only in testing or in predictions (for unassociated sources), but not in training. If the value is infinity, then we formally substitute it by the largest value found in the sample multiplied by 10. An unphysical zero (e.g., in significance) is substituted by the smallest value in the sample divided by 10, while a missing value is substituted by the average of the sample. There can be other ways to replace the missing or unphysical values, e.g., by using k nearest neighbors regression, but since the number of such sources is relatively small (13 for 3FGL and 14 for 4FGL-DR2), the choice of the method to replace the missing values does not significantly affect the results. In the final probabilistic catalogs, we use a column "Missing_Values_Flag" to mark the sources with missing or unphysical values.

As an example of the construction of a probabilistic catalog, we use the 3FGL catalog. In this section we perform a two-class classification to separate PS into pulsars and AGNs. Thus for training and testing, we subselect the sources, which are associated to pulsars and AGNs. The three-class classification into pulsars, AGNs, and other sources is discussed in Section 5. After the training of the algorithms, we test the performance with the test sources and predict the classes of the unassociated sources. The general workflow will have the following steps:

1. Select data for training and testing.
2. Optimize algorithms using training datasets. We select meta-parameters of the algorithms by optimizing accuracy of classification and test for overfitting using the test datasets. In order to get stable results, we repeat the separation of the data into training and testing samples 100 times and average the accuracy.
3. Make predictions for unassociated point sources in the 3FGL catalog. We also apply the classification to associated sources, which we use for consistency checks.

As a result of the analysis in this section, we select meta-parameters for the four ML algorithms, which we then use in the following section to construct probabilistic catalogs based on the *Fermi*-LAT 3FGL and 4FGL-DR2 catalogs.

3.1. Data and feature selection

For training of the algorithms we use the associated sources without missing or unphysical values, which were classified as either AGNs (classification labels in the 3FGL catalog: agn, FSRQ, fsrq, BLL, bll, BCU, bcu, RDG, rdg, NLSY1, nlsy1, ssrq, and sey) or pulsars (classification labels in 3FGL: PSR, psr). There are 1905 such sources in the 3FGL catalog.

There are several tens of features of point sources quoted in the catalog, such as the position, photon and energy fluxes integrated in different energy bands, spectral parameters, variability index, as well as corresponding uncertainties. We took some of the main features and also added 4 hardness ratios defined as

$$HR_{ij} = \frac{EF_j - EF_i}{EF_j + EF_i}, \tag{2}$$

where EF_i is the energy flux in bin i and $j = i + 1$ (i.e., the bins are consecutive).

Spectral index is one of the most important characteristic of sources. Unfortunately in the 3FGL catalog, the definition of the spectral index is different for associated and unassociated sources. In particular, the gamma-ray flux of pulsars is described by a power-law with a (super)exponential cutoff $\propto E^{-\Gamma} e^{-(E/E_c)^b}$, where the "Spectral_Index" feature in the catalog is the parameter Γ. On the other hand, gamma-ray flux of unassociated sources with significant curvature is represented by the log-parabola function $\propto (E/E_0)^{-\alpha-\beta\ln(E/E_0)}$, where the "Spectral_Index" feature is the parameter α, i.e., the tilt in the spectrum at the pivot energy E_0 (which also varies for different sources). Since the "Spectral_Index" feature has different definitions for associated pulsars and for possible pulsars among unassociated sources, its use for training the algorithms to separate pulsars from AGNs is problematic. If one fits all spectra of sources in the catalog by a power-law function, then the corresponding indices of the power laws are represented by "PowerLaw_Index" feature in the catalog. This feature is defined uniformly for all associated and unassociated sources, i.e., it is safe to use for training. Unfortunately, the power-law function is not a good description of the gamma-ray flux from pulsars. Consequently, in the classification of the 3FGL sources we have constructed a new feature: the index at 500 MeV (denoted in the following as "500MeV_Index"), defined as minus the derivative of the log flux:

$$n(500\,\mathrm{MeV}) = -\left.\frac{d\ln F}{d\ln E}\right|_{E=500\,\mathrm{MeV}} \tag{3}$$

For log-parabola and for power-law with (super)exponential cutoff it is respectively

$$n(500\,\mathrm{MeV}) = \alpha + 2\beta\ln(500\,\mathrm{MeV}/\mathrm{E}_0) \tag{4}$$

$$n(500\,\mathrm{MeV}) = \Gamma + b\,(500\,\mathrm{MeV}/E_c)^b \tag{5}$$

This feature has a more uniform definition for all sources in the 3FGL catalog than the Spectral_Index. It also has a better separating power than PowerLaw_Index, provided that pulsars have typically harder spectra at energies below 1 GeV than AGNs.

In order to select independent features, we calculate the Pearson correlation coefficients for the features: GLON (Galactic longitude), GLAT (Galactic Latitude), Sign_Avg (Signif Average), Pivote_E (Pivot Energy), FD (Flux Density), Un_FD (Uncertainty on Flux Density), F1000 (Flux_1000), Un_F1000 (Uncertainty_Flux_1000), E_F100 (Energy_Flux100), Un_E_F100 (Uncertainty_Energy_F100), Sig_Cur (Signif_Curvature), Sp_Ind (Spectral Index), Un_Sp_Ind (Uncertainty_Spectral_Index), PL_Ind (PowerLaw_Index), Var_Ind (Variability_Index), 500_Ind (Index at 500 MeV), HR_{ij} for the Hardness Ratios defined above. A graphical representation of the correlations is shown in Fig. 1. In the following, if two features have (anti)correlation $\gtrsim 0.75$ ($\lesssim -0.75$), then we keep only one of the features for classification. Taking into account the correlation among the features and the above discussion of the spectral index definition, we have selected the following eleven features for the classification of the 3FGL sources: Galactic latitude (GLAT), Galactic longitude (GLON), ln(Energy_Flux_100), ln(Unc_Energy_Flux100), 500MeV_Index, ln(Signif_Curve), ln(Variability_Index), and the four hardness ratios HR_{ij}. The table of features and their statistics can be found in Appendix A.

3.2. Construction of classification algorithms

The number of tunable parameters in the classification algorithms is not fixed a priori. Moreover, there is a certain freedom in the choice of the architecture of the algorithms, such as the number of hidden layers and the number of neurons in neural networks. In general, one starts with a simple model and increases the complexity (the number of tunable parameters) until the model can describe the data well, but does not overfit it. The overfitting is tested by splitting the input data into training and testing samples. The training sample is used for optimizing the parameters, while the test sample is used to check that the model is not overtrained (for overtrained models the accuracy on the test sample is significantly worse than the performance on the training sample). For our catalogs we split the data randomly into 70% training and 30% testing samples.

In this paper we determine the probabilistic classification of a source with an algorithm by the class with the maximal probability (as estimated by this algorithm). In the case of two classes, this is the class with probability larger than 0.5. In the case of three classes, the largest probability can be smaller than 0.5 but always larger than 1/3. Although the classification probabilities for some sources are not very large, e.g., a significant fraction of sources classified as pulsars may turn out to be AGNs or other sources, the main goal of our analysis is not to determine a list of sources, which are classified as pulsars or AGNs with high probabilities, but to determine the probabilities themselves and to estimate the uncertainties on the probabilities. In other words, our main goal is the construction of the probabilistic catalogs, which we make available online (SOM 2021). A user of these probabilistic catalogs can choose a smaller or a larger probability threshold for a particular class depending on the purpose of their analysis.

3.2.1. Random Forests

The two main parameters characterizing the RF algorithm are the number of trees and the maximum depth allowed in the trees. We use the Gini index as the objective function for the optimization of parameters (split values of features in the nodes).

Fig. 2 shows the dependence of the accuracy of the test sample as a function of maximum depth and the number of trees. The results for each point are averaged over 100 realizations of the split into training and testing samples. We notice that the accuracy does not decrease as the maximal depth of the trees increases, i.e., there is no overfitting as the complexity of the model increases with increased maximum depth.

This is due to the random choice of a subset of features at each node (maximal number of allowed features is $\sqrt{\#\text{ features}}$). It is also insensitive to the number of trees above approximately 20 trees. For classification we use 50 trees with a maximum depth of 6.

In order to illustrate the separation of PS into AGNs and pulsars, we retrain the RF algorithm using only two features: log of curvature significance and log of the variability index, and plot the resulting probabilities of classes in Fig. 3 for the model with 50 trees with a maximum depth of 6. The probabilities are averaged over 100 splits into training and testing samples. It is important to note that in this plot the model is trained on only two features. Nevertheless a good testing accuracy of 97% is reached, which is similar to the accuracy of the RF classification with all 11 features. For the final classification with RF, we use 11 features and average over 1000 splits into training and testing samples.

3.2.2. Boosted Decision Trees

The meta-parameters for BDT algorithms are similar to RF algorithms: the number of trees and the maximal

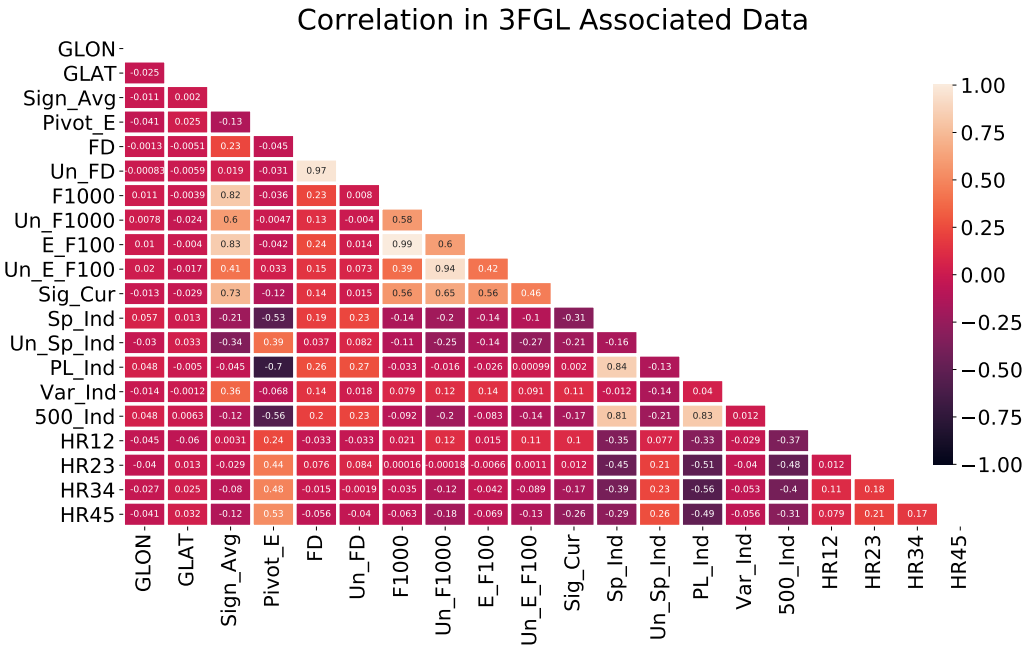

Fig. 1. The correlation matrix of features for the associated sources in the 3FGL catalog. The "500MeV_Index" is defined in Eq. (3). The hardness ratios (HR12, HR23, etc.) are defined in Eq. (2). All other features are taken directly from the 3FGL catalog. See text for the description of labels.

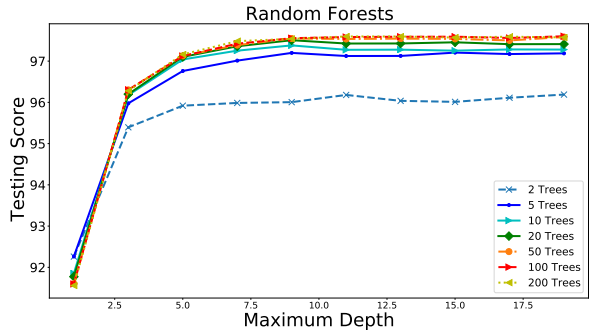

Fig. 2. Test score (accuracy) of RF classification as a function of the number of trees and the maximal depth of trees.

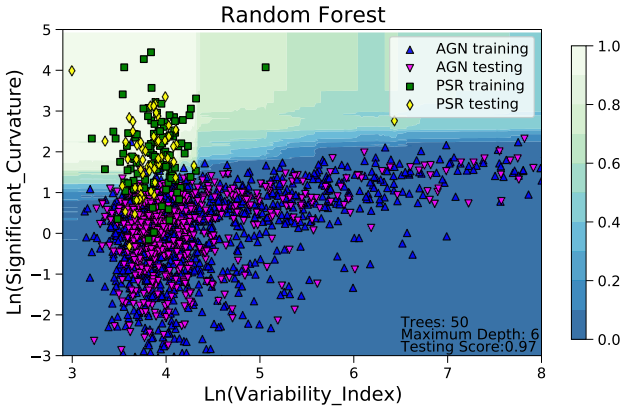

Fig. 3. RF classification domains showing class probabilities for training with two features averaged over 100 random splits into training and testing samples. One of these splits is shown for illustration. Color scale describes the probability for a source to be a pulsar.

depth. We used the Gradient Boosting algorithm for the construction of BDT (Friedman 2001b). The classification is performed by a weighted average of trees, where the trees are constructed recursively in order to better address misclassifications from the previous step. Dependence of the accuracy on tree depth is shown in Fig. 4. Unlike the RF, which is also an ensemble based method, the testing accuracy drops for the maximal depths larger than 7.

The classification domains in case of two features for 20 trees and the maximum depth of 2 is presented in Fig. 5. For the classification we will use BDT with 100 trees and the maximum depth of 2.

In tree-based algorithms, one can calculate feature importance by using the averaged reduction of impurity for nodes (Gini index in our case) involving the different features. The importance of features for the case of two different algorithms: RF with 50 trees and maximum depth of 6, and BDT with 100 trees with maximum depth of 2, are shown in Table 1. We find that the most important feature for both cases is the significance of curvature. Other significant features are the hardness ratio of the last two energy bins, uncertainty of the energy flux at 100 MeV, and the variability index.

It is interesting to note that Galactic latitude is among the least significant features. We have also used

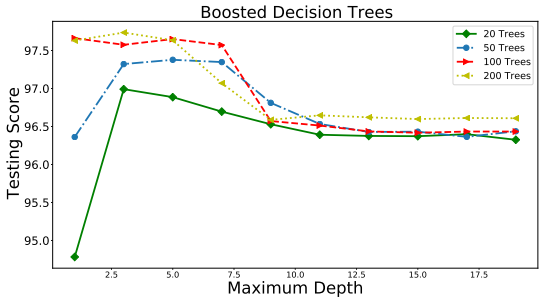

Fig. 4. Dependence of BDT accuracy on maximum depth and the numbers of trees.

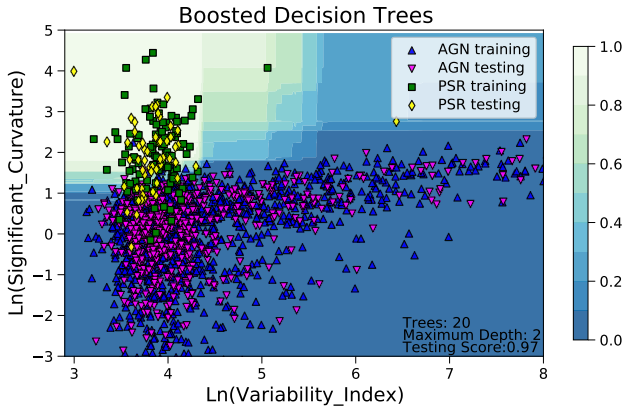

Fig. 5. Classification domains for BDT for training with two features averaged over 100 splits into training and testing samples.

Feature	RF: 50, 6	BDT: 100, 2
ln(Signif_Curve)	0.331	0.518
HR45	0.137	0.071
ln(Unc_Energy_Flux100)	0.122	0.050
ln(Variability_Index)	0.098	0.225
ln(Energy_Flux100)	0.071	0.019
500MeV_Index	0.065	0.028
HR23	0.062	0.052
HR12	0.052	0.012
HR34	0.025	0.005
GLAT	0.017	0.002
GLON	0.014	0.011

Table 1. Feature importances for RF (50 trees, max depth 6) and BDT (100 trees, max depth 2) algorithms. The features are ordered by decreasing importance in the case of the RF algorithm.

sin(GLAT) to check that this is not due to scaling, i.e., the large range of values of GLAT, but the significance is similar to the GLAT itself. We further discuss the dependence on GLAT in Section 6.2, where we calculate the latitude and longitude profiles of the associated and unassociated source counts.[1]

[1] Feature importances for the classification of 4FGL-DR2 sources with RF and BDT algorithms are reported in Appendix A.

3.2.3. Neural Networks

In the case of NN, the number of free parameters depends on the number of hidden layers and on the number of neurons in the hidden layers. The final model accuracy also depends on the number of epochs that the network is allowed to be trained for and on the optimization algorithm.

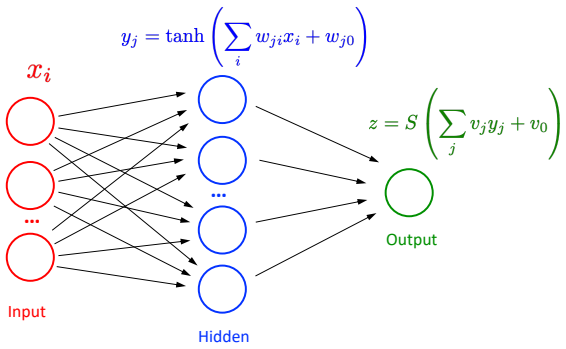

Fig. 6. NN architecture that we use in the construction of the probabilistic catalogs. The activation function in the output layer is sigmoid $S(x) = e^x/(1+e^x)$.

The general architecture of the NN that we use in this paper is shown in Fig. 6. It is a fully connected NN with 11 input nodes (shown by red circles with input features x_i), one hidden layer (shown by blue circles), and an output layer (shown by the green circle). The hidden layer consists of several nodes with values y_j. For the activation function at the hidden layer we use either hyperbolic tangent (tanh - shown on the plot) or rectified linear unit (relu). The activation function for the output layer is sigmoid, which we use to make sure that the output value can be interpreted as a class probability. The unknown parameters are weights of features in the hidden layer w_{ji} and in the output layer v_j including offsets w_{j0} and v_0. The unknown parameters are optimized by minimizing a loss function, which we choose to be the cross entropy $-\log L = -\sum_i (y_i \log(p_i) + (1-y_i)\log(1-p_i))$, where $y_i = 0, 1$ are the true labels of the sources and p_i are the predicted class probabilities. We have also used NN with two hidden layers, but the accuracy was similar to the networks with one hidden layer (Appendix A). For the final classification model, we have chosen to use one hidden layer.

Dependence of the testing accuracy on the number of neurons in the hidden layer, on the activation function, and on the optimization algorithm is shown in Fig. 7. We compare two activation functions at the hidden layer (tanh and relu) and two optimization algorithms: Limited memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS, Liu & Nocedal 1989) and the stochastic gradient descent algorithm Adam (Kingma & Ba 2014). We use 300 epochs for training. Around 11 neurons in

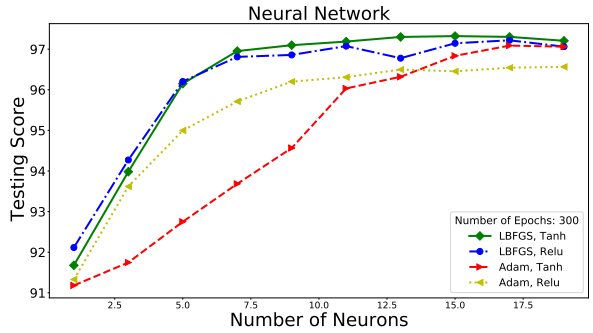

Fig. 7. Dependence of accuracy on the number of neurons for different NN models.

the hidden layer appears to be an optimal choice, since increasing the number of neurons leads to no significant increase in accuracy for all models.

Dependence on the number of epochs (number of iterations in fitting) is presented in Fig. 8. The accuracy increases with higher number of epochs and saturates at around 200 for LBFGS and 300 for Adam.

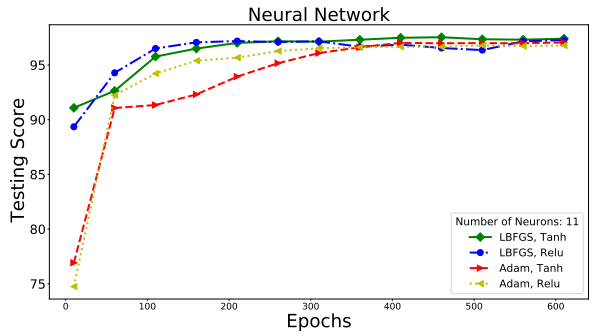

Fig. 8. Dependence of testing accuracy on the number of epochs in training for different solvers and activation functions.

We illustrate the classification domains for NN with two input features in Fig. 9. In this case we also use only two neurons in the hidden layer. One can see that the separation boundary is smoother compared to the RF domains in Fig. 3 or BDT domains in Fig. 5. For our final model we chose one hidden layer with eleven neurons, 300 training epochs, LBFGS solver, and tanh activation function at the hidden layer.

3.2.4. Logistic Regression

As we have discussed in Section 2.2, the probability to belong to class 1 or 0 in LR is represented by the sigmoid function $p_1(x) = 1 - p_0(x) = \frac{e^{m(x)}}{1+e^{m(x)}}$ (see Eq. (1)), where $m(x)$ is a function of input features x. The complexity of the model is given by the number of parameters in $m(x)$. We have considered two cases for $m(x)$: linear and quadratic function of the input features x. Quadratic $m(x)$ resulted in a similar accuracy as linear $m(x)$. Consequently, we have restricted our attention to linear functions $m(x) = f_0 + \sum_{k=1}^{11} f_k x_k$. In Fig. 10 we show the accuracy of the LR method as a function of the number of iterations for different solvers, e.g., LBFGS (Liu & Nocedal 1989), Stochastic Average Gradient (SAG, Schmidt et al. 2017), SAGA (a variant of SAG, Defazio et al. 2014), and liblinear (a special solver for LR and support vector machine classifications, Fan et al. 2008). As one can see from Fig. 10, LBFGS and Liblinear outperform the other two solvers and converge much faster. In order to illustrate the probability domains in LR, we show the classification with two features (LBFGs, 200 iterations) in Fig. 11. The domains look similar to the domains in the NN case (Fig. 9). For the final classification we will use LBFGs solver with 200 iterations.

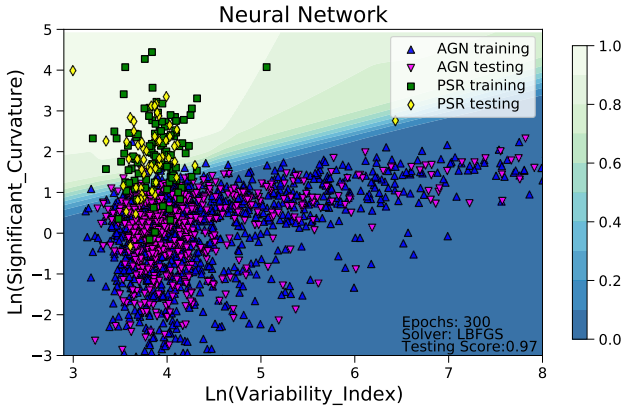

Fig. 9. NN classification domains for 2 input features averaged over 100 random splits into training and testing samples. We use 2 neurons in the hidden layer, tanh activation function, and LBFGS solver.

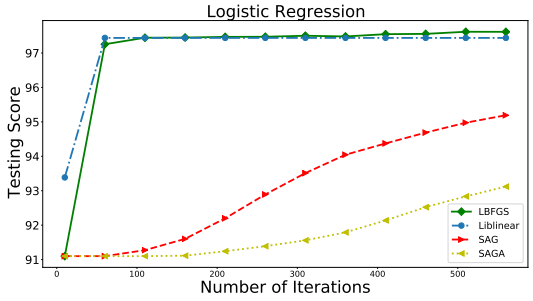

Fig. 10. Dependence of LR testing accuracy on the number of iterations for different solvers.

3.3. Oversampling

Fermi-LAT catalogs have many more AGNs than pulsars, i.e., the datasets are imbalanced. For example, the 3FGL catalog has 1744 associated AGNs (1739 without missing or unphysical values) and 167 associated pulsars (166 without missing or unphysical values). In the previous subsections we have optimized the overall accuracy.

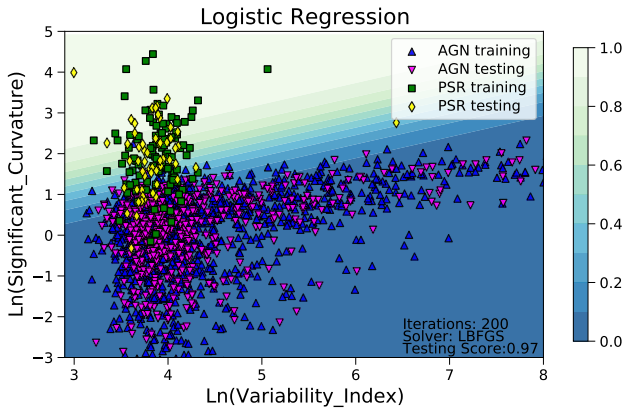

Fig. 11. Classification domains for LR with two features averaged over 100 random splits into training and testing samples.

In this case, the algorithms try to identify AGNs rather than pulsars, since it gives better accuracy. As a result, in the region of parameter space, where both pulsars and AGNs are present, the algorithms will give higher probability for a source to be an AGN.

The problem of classification of imbalanced datasets can be quantitatively described in terms of precision and recall. If we denote by "# true" the number of pulsars in the dataset, by "# positive" – the number of sources predicted to be pulsars, and by "# true positive" – the number of pulsars predicted to be pulsars, then $precision = \frac{\#\text{ true positive}}{\#\text{ positive}}$ is a measure how clean the prediction is, while $recall = \frac{\#\text{ true positive}}{\#\text{ true}}$ is a measure how well the algorithm can detect pulsars, i.e., how complete the list of predicted pulsars is. If we reduce the pulsar domain by attributing uncertain sources predominantly to AGNs, then for pulsars the precision will increase, but the recall will decrease.

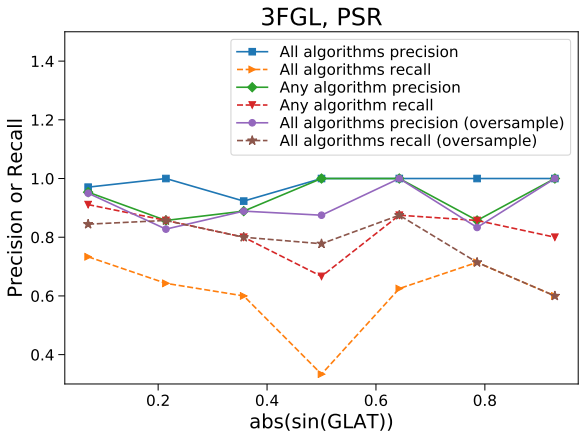

Fig. 12. Precision and recall for pulsars using all-algorithm and any-algorithm classification for unweighted training data and all-algorithm classification for oversampling of pulsars in training data. For details see Section 3.3.

In Fig. 12 we show precision and recall for classification of pulsars. In particular, in the first two lines (solid blue with squares and dashed orange with right triangles) a source is categorized as a pulsar if all four algorithms classify it as a pulsar, while in lines 3 and 4 (solid green with diamonds and dashed red with down triangles) a source is attributed to the PSR class, if any of the algorithms classifies it as a pulsar. It is clear that for lines 1 and 2 the pulsar domain is smaller than for lines 3 and 4, since in the former case, the domain is the intersection of domains for individual algorithms, while in the latter it is the union. For all-algorithms classification the precision is 100% for most of latitudes, while the recall is between 40% and 80%, i.e., the list of pulsars is generally clean but incomplete. In case of any-algorithm classification, the recall is increased by about 20% for most latitudes compared to the all-algorithms classification, but the precision drops by up to 20% at some latitudes, i.e., the completeness improves at the expense of cleanliness of the sample. Alternatively to using any-algorithm classification, one can give larger weights to pulsars or oversample pulsars in the training process, i.e., use the same source several times, so that the numbers of pulsars and AGNs in training are the same. Provided that in some applications it is beneficial to have as complete a list as possible of pulsar candidates among unassociated sources, we have retrained the algorithms using oversampling with the same meta-parameters as in the previous sections.

In general one can either under- or oversample a dataset. Undersampling would reduce the number of AGNs to match the number of pulsars. However, since the total number of sources is not very high, we chose to oversample the data. For training with oversampling, we copy randomly existing pulsars and add them to the dataset until the number of pulsars and AGNs are the same. Although pulsars in the training dataset are redundant, they help to increase the weight of pulsars in the classification model. We illustrate the oversampling procedure in Fig. 13 top panel: the number of times a source appears in training is shown by adding markers with shifts to the right and above the original position of the source (note that the shift is introduced for presentation only, the parameters of the sources are exactly the same as in the original source). In the bottom panel of Fig. 13 we repeat Fig. 11 in order to compare the classification domains with and without oversampling. One can see that pulsar domain in the top panel is larger than the pulsar domain in the bottom panel. As a result, in the top panel more pulsars are classified as pulsars but also more AGNs are falsely classified as pulsars in the intersection region. Since the overall number of AGNs is

larger than the number of pulsars, the testing accuracy with oversampling is smaller than without oversampling.

The results of training with oversampling are presented in Fig. 12, lines 5 and 6 (solid purple with circles and dashed brown with stars). These lines show precision and recall when a source is categorized as a pulsar, if all four algorithms trained with oversampling classify it as a pulsar. The precision and recall in this case are similar to the any-algorithm classification for the training without oversampling.

In order to test the oversampling method, we compare in Appendix B the oversampling-by-repetition with Synthetic Minority Over-sampling Technique (SMOTE, Chawla et al. 2011). The result of the comparison is that for class probabilities of individual sources, the difference in oversampling is generally smaller than the uncertainty due to the random choice of the training sample, while the differences in population studies are comparable to the differences among the different algorithms.

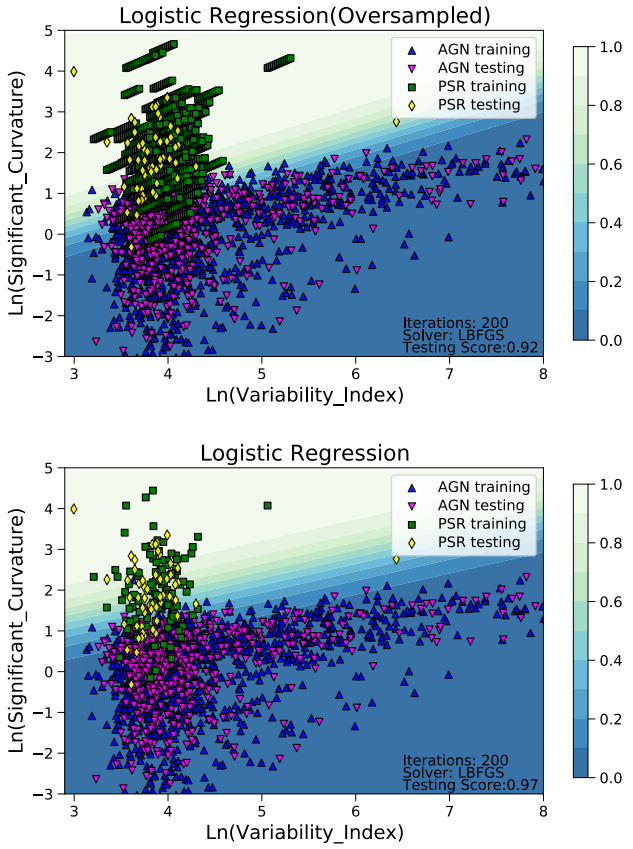

Fig. 13. Top panel: LR classification domains showing class probabilities for training with oversampling. The oversampling is illustrated by repeating the pulsar markers with a shift: the number of markers is equal to the number of times the pulsar appears in training. Bottom panel: we repeat Fig. 11 for convenience of comparison with the oversampled training in the top panel. In both panels the domains are obtained by averaging over 100 random splits into training and testing samples.

4. Probabilistic catalogs based on the 3FGL and 4FGL-DR2 catalogs

In this section we use the ML algorithms optimized in the previous section to construct a probabilistic classification of sources in the 3FGL and 4FGL-DR2 catalogs.

4.1. Probabilistic classification of sources in 3FGL and comparison with 4FGL-DR2

We use the following four algorithms for the classification of sources: RF with 50 trees and maximal depth of 6, BDT with 100 trees and maximal depth of 2, NN with 11 neurons, LBFGS solver, and 300 epochs, and LR with LBFGS solver and 200 iterations. For training we use the pulsars and AGNs from the 3FGL catalog without missing or unphysical values. In addition to original datasets, we perform oversampling of pulsars in order to balance the numbers of pulsars and AGNs. As a result, we have 8 classification methods: 4 algorithms trained with and without oversampling.

Algorithm	Parameters	Testing Accuracy	Std. Dev.	Comparison with 4FGL-DR2 Accuracy
RF	50 trees, max depth 6	97.37	0.60	91.09
RF_O		97.90	0.50	89.44
BDT	100 trees, max depth 2	97.65	0.54	90.43
BDT_O		97.79	0.51	91.75
NN	300 epochs, 11 neurons, LBFGS	97.29	0.97	90.10
NN_O		94.31	5.13	87.13
LR	200 iterations, LBFGS solver	97.63	0.54	90.43
LR_O		93.68	0.99	85.15

Table 2. Testing accuracy of the 4 selected algorithms for classification of 3FGL sources and comparison with associations in the 4FGL-DR2 catalog. "_O" denotes training with oversampling.

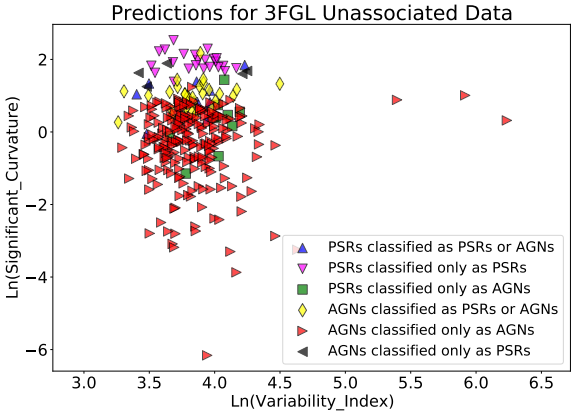

Fig. 14. Comparison of class prediction for unassociated 3FGL sources with classes in 4FGL-DR2. For more details see Section 4.1.

The selected algorithms are summarized in Table 2, where oversampling is shown by "_O". "Average testing accuracy" is computed by taking 1000 times the 70% - 30% split into training and testing samples and averaging over the accuracies computed for the testing samples. In addition, we look at sources, which are unassociated in 3FGL but have either pulsar or AGN association in 4FGL-DR2: there are 303 such sources. The accuracy of our prediction for the four selected algorithms with and without oversampling, taking the 4FGL-DR2 classes as the true values, is reported in the column "Comparison with 4FGL Accuracy".

As a result of the classification with the eight ML methods, we created a probabilistic catalog based on the 3FGL sources. We train on 70% of the sources associated with pulsars or AGNs without missing or unphysical values (there are thirteen sources with missing or unphysical values in the 3FGL catalog: 2 unassociated, 5 AGNs, 1 pulsar, and 5 "other" sources). We replace the missing and unphysical values according to the procedure described at the beginning of Section 3. We calculate the probabilities of classes for testing sources, for sources which are not classified as pulsars or AGNs or have missing or unphysical values, and for unassociated sources. We repeat the splitting and training 1000 times and report the sample average and standard deviation of the classification probabilities, i.e., we average over 1000 values for unassociated sources, sources not classified as AGNs or pulsars, and sources with missing or unphysical values, while the average for AGNs and pulsar without missing or unphysical values is over the number of times the sources appear in the testing sample, which is 300 on average.

In the probabilistic catalogs we add columns with corresponding probabilities for each algorithm and each class, i.e., provided that there are 8 methods (including oversampling) and 2 classes, we add 16 columns: 8 for unweighted and 8 for oversampled training data. The columns with '_O' represent the oversampled probabilities. We also add 16 columns for the standard deviations of probabilities. Although class probabilities and standard deviations for each algorithm are not independent (probabilities add up to 1 and standard deviations are equal for AGN and PSR classes), we keep the corresponding columns in view of the generalizations to multi-class classification (e.g., the 3-class classification in Section 5).

Table 3 shows an example of the probabilistic catalog for a few unassociated 3FGL sources. Notice that the last source is classified as a pulsar by BDT and RF algorithms and as an AGN by LR and NN algorithms. It is therefore an example of a source with mixed classification.

For the determination of candidate classes based on the probabilistic classification, we consider the following two conditions: 1) that all algorithms agree, i.e., each algorithm predicts the same class for a source with more than 50% probability,[2] and 2) that the sum of probabilities for a source to belong to a certain class is larger than 7 (this means that on average the probability is larger than $7/8 = 0.875$). Both of these conditions are stricter than the classification using probabilities for any of the 8 algorithms. For convenience, we add a column in the probabilistic catalogs with the most likely probabilistic classes of sources based on the condition that all algorithms agree on the classification (sources with mixed classification are labeled as "MIXED" in this column). In order to test the performance of the classification conditions we compare the precision and recall for unassociated source in the 3FGL catalog, which are associated in the 4FGL-DR2, and also calculate the expected precision and recall based on test samples used in training. In total there are 340 sources, which are unassociated in 3FGL but have associations in 4FGL-DR2. The result of classification of the unassociated sources in the 3FGL catalog, which have associations in the 4FGL-DR2 catalog, using the condition that all methods agree on the classification are presented in Table 4. "MIXED" column shows the numbers of sources, for which different algorithms predict different classifications. Columns show the predictions for the 3FGL unassociated sources, while the rows show the associations in the 4FGL-DR2 catalog. For the later comparison with the 3-class classification, we also add unassociated source in the 3FGL

	AGN Probability			
Source_Name_3FGL	BDT	RF	LR	NN
3FGL J0000.2-3738	1	0.995	0.991	0.993
3FGL J0002.0-6722	1	0.996	0.996	0.994
3FGL J0002.6+6218	0.005	0.05	0.054	0.046
3FGL J0506.3-0357	0.569	0.653	0.378	0.227

Table 3. Example of the AGN classification probabilities for a few unassociated sources in the 3FGL catalog (Acero et al. 2015). We have omitted the oversampled probability columns here.

[2] We use the 50% threshold in the 2-class case for illustration. Even if the probabilities for a source to be, e.g., a pulsar is larger than 50% for all eight methods, there is still a large chance for the source to be an AGN if the probabilities are around 50%. For this reason, the classes of sources reported in this work derived from the class probabilities should be viewed as candidate classes. Depending on the application, a higher (lower) threshold can be used for a cleaner (more complete) sample. We study the effect of changing the classification threshold on precision and recall for the all-algorithms-agree method in Appendix C. The full catalogs with all probabilities are available online (SOM 2021). More confident classification of particular sources can be obtained with multiwavelength studies, which are beyond the analysis in this paper.

catalog, which have associations with sources other than pulsars and AGNs.

We also present in Table 4 the precision and recall estimates from the comparison of the 3FGL and 4FGL-DR2 catalogs ("4FGL assoc" columns and rows). For example, the precision for AGNs is the number of true positive predictions (223) over the number of positive predictions (the sum of numbers in the AGN column: (223 + 10 + 8), which gives $223/241 \approx 0.93$. The estimated precision for PSRs is $23/(23+5+6) \approx 0.68$. In addition we show the expected precision and recall using the condition that all methods agree ("all agree" columns and rows) and the condition that the sum of probabilities is larger than 7 ("$\sum_a p_a > 7$" columns and rows) calculated using the class probabilities for associated sources reported in the probabilistic catalogs (these probabilities are derived as an average over the test samples). The precision (recall) estimates using, e.g., the condition that all algorithms agree is computed by taking the ratio of associated pulsars in 3FGL, which are also predicted to be pulsars by all 8 algorithms, to the number of sources associated with pulsars (to the total number of sources predicted to be pulsars among associated sources) in the 3FGL catalog. The precision and recall estimates for the $\sum_a p_a > 7$ condition are computed analogously. The latter condition is on average stricter than the "all agree" condition, which results in larger precision and smaller recall than in the "all agree" case. Also all unassociated 3FGL pulsar candidates, which satisfy $\sum_a p_a > 7$ condition, are associated to pulsars in the 4FGL-DR2.

We note that the precision and recall calculated with the "all agree" condition from the comparison of 3FGL and 4FGL-DR2 catalogs ~~is worse~~ are smaller than the precision and recall ~~expectations using the estimates from the test samples in training~~determined using the test samples and the "all agree" and $\sum_a p_a > 7$ conditions (Table 4). In order to understand the reason for the worse performance of classification in comparison of 3FGL predictions with the 4FGL-DR2 associations relative to the expectations, we plot in Fig. 14 the 303 sources unassociated in 3FGL but with PSR or AGN associations in 4FGL-DR2. The class at the beginning of the label name in Fig. 14 corresponds to the association in the 4FGL-DR2, while the second half of the labels corresponds to classification of unassociated sources in 3FGL. For example, "PSRs classified only as PSRs" shows sources which have a PSR association in 4FGL-DR2 and all eight methods classified the corresponding unassociated sources in 3FGL as a pulsar. "PSRs classified as either PSRs or AGNs" labels sources with PSR associations in 4FGL-DR2 but the corresponding unassociated sources in 3FGL have both PSR and AGN classifications by different ML methods. We notice that misclassified or partially misclassified sources in Fig. 14 typically happen on the boundary between the two classes or even inside the opposite class. Many of these sources also have flags in the 3FGL catalog, such as a potential problem with the background diffuse emission model in the location of the source, which can lead to a poor reconstruction of the source spectrum and, consequently, misclassification of the source. ~~The estimated precision for AGNs (pulsars) using either "all-agree" or $\sum_a p_a > 7$ condition is better than about 97% (80%) with the 2-class classification, while the precision estimates~~ Thus, we find that the precision and recall calculated from the comparison of 3FGL and 4FGL-DR2 catalogs ~~is about 93% (70%) for AGNs (pulsars). We will see in Section 5 that the expected precision using either "all-agree" or $\sum_a p_a > 7$ condition for the 3-class classification for AGNs (pulsars) is better than about 99% (90%), but the precision estimated from the comparison of 3FGL and 4FGL-DR2 catalogs is still about 93% (70%). Thus even if the expected precision is better in the 3-class case than in the 2-class case, the precision estimates from the comparison of 3FGL~~ give more realistic estimates of the true precision and ~~4FGL-DR2 catalogs are similar in the 2- and 3-class cases. We conclude that the smaller precision in the comparison of the 3FGL and 4FGL-DR2 catalogs is mostly due to errors and uncertainties in the input data and represents an irreducible error of the analysis~~recall compared to the estimates determined from the test samples, since the former take into account possible errors in the reconstruction of source parameters (such as the spectral parameters) as well as the differences in distributions of associated and unassociated sources.

4FGL-DR2 class	3FGL prediction			Recall	Recall	Recall
	AGN	PSR	MIXED	(4FGL assoc)	(all agree)	($\sum_a p_a > 7$)
AGN	223	5	30	0.86	0.94	0.92
PSR	10	23	12	0.51	0.67	0.61
OTHER	8	6	23	–	–	–
Precision (4FGL assoc)	0.93	0.68	–			
Precision (all agree)	0.97	0.78	–			
Precision ($\sum_a p_a > 7$)	0.98	0.82	–			

Table 4. Comparison of classes predicted for unassociated sources in the 3FGL catalog using 2-class classification with associations in the 4FGL-DR2 catalog. The precision and recall are calculated using all-algorithms-agreement condition and 4FGL-DR2 associations ("4FGL assoc" labels) or the class probabilities of associated 3FGL sources derived from test samples using either all-algorithms-agreement condition ("all agree" labels) or the $\sum_a p_a > 7$ condition ("$\sum_a p_a > 7$" labels). The AGN and PSR sources are also represented in Fig. 14.

We summarize the results of classification of unassociated 3FGL sources with the 2-class classification in Table 5 in the 3FGL "2-class" row. The "AGNs" column

Catalog	Classification	AGN	PSR	OTHER	MIXED
3FGL	2-class	599	111	–	300
	2-class corr	580.0	97.0	56.4	276.5
	3-class	587	53	69	301
4FGL-DR2	2-class	878	162	–	627
	2-class corr	826.2	134.5	140.4	565.9
	3-class	739	64	274	590

Table 5. Expected number of AGNs, pulsars, and other sources as well as sources with mixed classifications among the unassociated 3FGL and 4FGL-DR2 sources derived with the 2-class (Section 4) and 3-class (Section 5) classification. The "2-class corr" row shows correction of the 2-class classification prediction due to the presence of OTHER sources among the unassociated ones (see Section 4.1 for details).

shows the number of unassociated sources predicted to be AGNs using the all-algorithms-agree condition. Similarly the "Pulsars" column shows the number of unassociated sources where all the algorithms predict the source to more likely be a pulsar. The "Mixed" column shows the number of sources with mixed classification, i.e., some algorithms predict that the source is more likely an AGN while the other algorithms predict that it is more likely a pulsar. We also add the "OTHER" column in order to compare the results with the 3-class classification in Section 5. Since there is no "OTHER" class in the 2-class classification, the corresponding entry is empty. Out of 1010 unassociated sources in 3FGL, 111 are classified as pulsars by all eight methods, 599 are classified as AGNs, and 300 have mixed classifications.

In the "2-class corr" row of Table 5 we show a possible correction of the number of pulsars and AGNs due to the presence of other sources. Here we assume that the fraction of AGN-like and pulsar-like sources among the other sources is the same for associated and for unassociated sources. In particular, if we denote by $N_{\rm AGN}$ the number of unassociated sources with AGN-like probabilistic classification, by $N_{\rm AGN}^{\rm ass\,OTHER}$ the number of sources with AGN-like classification among associated OTHER sources, by $N_{\rm ass}$ ($N_{\rm unass}$) the total number of associated (unassociated) sources, then the number of AGN-like sources among the unassociated ones corrected for the presence of OTHER sources can be estimated as

$$N_{\rm AGN}^{\rm corr} = N_{\rm AGN} - N_{\rm AGN}^{\rm ass\,OTHER}\,\frac{N_{\rm unass}}{N_{\rm ass}}. \quad (6)$$

Analogous corrections are applied for the number of unassociated sources with PSR and with mixed classifications. If we denote by $N^{\rm ass\,OTHER}$ the total number of associated other sources, then the estimated number of OTHER sources among unassociated ones is

$$N_{\rm OTHER}^{\rm unass} = N^{\rm ass\,OTHER}\,\frac{N_{\rm unass}}{N_{\rm ass}}. \quad (7)$$

We show this estimate in the OTHER column in the "2-class corr" row. We note that since $N_{\rm AGN}^{\rm ass\,OTHER} + N_{\rm PSR}^{\rm ass\,OTHER} + N_{\rm MIXED}^{\rm ass\,OTHER} = N^{\rm ass\,OTHER}$, this estimate is consistent with corrections in Eq. (6) for sources classified as AGNs, pulsars, or with mixed classification. The expected numbers of sources for the 4FGL-DR2 catalog and in the 3-class case are calculated in Section 4.2 and in Section 5 respectively.

4.2. Probabilistic classification of sources in the 4FGL-DR2 catalog

In this section we construct a probabilistic classification of sources in the 4FGL-DR2 catalog. The 4FGL-DR2 catalog (Ballet et al. 2020) is based on 10 years of *Fermi*-LAT data (compared to 8 years of data in the 4FGL catalog, Abdollahi et al. 2020). It contains 5788 sources, which is 723 sources more than in the 4FGL catalog (all sources in 4FGL are kept in 4FGL-DR2 even if they fall below the detection threshold with 10 years of data). In the 4FGL-DR2 catalog, 3503 sources are associated to AGNs, 271 sources are associated to pulsars, 1658 sources are unassociated (we only look at CLASS1 column in the catalog), and the rest 346 sources are other sources, such as PWN, SNR, etc. There are 14 sources in 4FGL-DR2 with missing or unphysical values: four AGNs, one PWN (Crab), and nine unassociated sources. As in the previous section, we use sources associated with either AGNs or pulsars for training, which have no missing or unphysical values. The unphysical and missing values are replaced according to the procedure described at the beginning of Section 3. We calculate the classification probabilities of AGN and PSR classes for both the associated and the unassociated sources.

The 4FGL-DR2 catalog has a higher number of features, especially due to the difference in modeling of the spectra compared with the 3FGL catalog. We selected 28 of these features plus 6 hardness ratios HR12, ..., HR67 (the 4FGL-DR2 catalog has 7 energy bins) and looked for correlations among them.

If any feature was correlated or anti-correlated with a Pearson index of ± 0.75 or higher with another feature, then only one of these features was kept. The resulting 16 features are: GLON, GLAT, ln(Pivot_Energy), ln(Energy_Flux100), ln(Unc_Energy_Flux100), LP_Index, Unc_LP_Index, LP_beta, LP_SigCurv, ln(Variability_Index), and the 6 hardness ratios.

For the classification of 4FGL-DR2 sources, we confirmed that the parameters used in 3FGL classification provide an optimal performance also for the 4FGL-DR2 catalog, except for NN, which requires more neurons in the hidden layer in the 4FGL-DR2 case. Therefore, we used the same meta-parameters for the four algorithms

as in the construction of the probabilistic catalog based on 3FGL, except for NN where we increased the number of neurons in the hidden layer to 16. Similar to the construction of the 3FGL probabilistic catalog, we use both unweighted training samples and oversampling, i.e., we have 8 classification methods. We retrain the algorithms using the 16 features for the 4FGL-DR2 sources. The corresponding accuracies are reported in Table 6.

Algorithm	Parameters	Testing Accuracy	Std. Dev.
RF	50 trees, max depth 6	97.87	0.36
RF_O		97.56	0.39
BDT	100 trees, max depth 2	97.63	0.39
BDT_O		97.72	0.38
NN	300 epochs, 16 neurons, LBFGS	97.41	0.47
NN_O		95.48	0.66
LR	200 iterations, LBFGS solver	97.80	0.38
LR_O		96.03	0.53

Table 6. Testing accuracy of the 4 algorithms on 4FGL-DR2 associated data. "_O" denotes training with oversampling.

	classification	AGN	PSR	OTHER
2-class precision	all agree	0.96	0.71	–
	$\sum_a p_a > 7$	0.97	0.78	–
2-class recall	all agree	0.95	0.70	–
	$\sum_a p_a > 7$	0.94	0.64	–
3-class precision	all agree	0.98	0.89	0.80
	$\sum_a p_a > 7$	0.99	0.94	0.85
3-class recall	all agree	0.94	0.62	0.38
	$\sum_a p_a > 7$	0.87	0.30	0.06

Table 7. The expected precision and recall for the classification of 4FGL-DR2 sources for two classification conditions: all algorithms agree ("all agree" rows) and that the sum of probabilities for the 8 algorithms is larger than 7 ("$\sum_a p_a > 7$" rows). For details on the 2- and 3-class classification see Sections 4.2 and 5 respectively.

Analogously to the 3FGL catalog, we use the agreement among the algorithms for the probabilistic classification of sources. We report the corresponding precision and recall in Table 7, where we also show precision and recall for the $\sum_a p_a > 7$ condition and for the 3-class classification described in Section 5. Using the agreement among the algorithms condition, we calculate the expected numbers of pulsars and AGNs among the 1658 unassociated sources in 4FGL-DR2 without missing values (see 4FGL-DR2 part of Table 5). The definition of rows is the same as in the 3FGL catalog 2-class classification in Section 4.1.

Finally, we looked at sources which were unassociated in both 3FGL and 4FGL-DR2 (using 'ASSOC_FGL' as an identifier for 3FGL sources). Out of 303 such sources[3], 40 sources are predicted to be pulsars using 3FGL features and 75 sources are predicted to be pulsars using 4FGL-DR2 features. This leads to 29 sources which are predicted by all eight methods to be pulsars for features taken from both the 3FGL and 4FGL-DR2 catalogs. For convenience, we save these 29 pulsar candidates as a separate file ("3FGL_4FGL-DR2_Candidates_PSR.csv" in the supplementary online materials (SOM 2021)). Out of these 29 sources classified as pulsars, 4 sources have counterparts in Parkes survey (Camilo et al. 2015) within 2 arc minutes (see Table 8). The data for the Parkes association candidates is also added to the "3FGL_4FGL-DR2_Candidates_PSR.csv" file.

Source_Name_4FGL	GLON	GLAT	Sep (arcsec)
4FGL J0933.8-6232	282.2	−7.9	52.9
4FGL J1539.4-3323	338.8	17.5	60.5
4FGL J1808.4-3358	358	−6.7	103.7
4FGL J2112.5-3043	14.9	−42.4	88.4

Table 8. Connection of unassociated 3FGL and 4FGL-DR2 sources classified as pulsars with Parkes pulsars (Camilo et al. 2015). GLON and GLAT are taken from 4FGL-DR2 and the separations in arcseconds with Parkes pulsars are given in the "Sep (arcsec)" column.

5. Three-Class Classification

One of the caveats of the analysis with two classes is that there are associated sources which do not belong to the AGN or PSR classes. These sources have the labels: unk, spp, glc, snr, gal, sbg, GAL, sfr, bin, SNR, HMB, LMB, css, PWN, pwn, hmb, SFR, BIN, lmb, NOV. We collect all associated sources, which do not belong to AGN and PSR classes, into a new class, which we label as "OTHER". Since in two-class classification we train algorithm to classify sources only into AGN and PSR classes, OTHER sources are also classified as either AGN or PSR. This introduces a bias in the estimates of the number of AGNs and pulsars among unassociated sources. One possibility to correct this bias is to assume that the fraction of OTHER sources among associated and unassociated sources are the same (Eq. (6)). This correction can be applied for the total number of sources or for the number of sources in some window of parameters, e.g., in a flux bin or in a range of latitudes and longitudes. This is a straightforward calculation but it has some limitations. In particular, it implicitly assigns one probability for each

[3] The 303 unassociated 4FGL-DR2 sources correspond to 302 unassociated 3FGL sources, because there are two 4FGL-DR2 sources, which are associated with one 3FGL source.

AGN and one probability for each PSR to belong to the OTHER class inside a selected range of parameters. For a small range of parameters the variance of this estimate can be very large due to a small number of associated OTHER sources in this parameter range. As we will see in Section 6, this correction depends on the choice of the variable used for binning, e.g., overall correction with latitude bins is not equal to the correction with longitude bins.

In this section we discuss the construction of probabilistic catalogs with multi-class classification (3-class classification in our case). We start with the construction of the probabilistic catalog based on 3FGL by adding the class "OTHER", which includes all associated sources without AGN or PSR associations: there are 113 such sources in 3FGL (108 OTHER sources are without missing or unphysical values). We use the same 11 features as in the 2-class classification: the only difference is that we use cos(GLON) instead of GLON. The reason for this is that the LR and NN methods have a significantly worse performance than RF and BDT methods when we use GLON, but, as we show below, all four methods have comparable accuracy when we use cos(GLON). This may be due to a discontinuity in the GLON variable. We perform optimization of the meta-parameters for the four ML algorithms with the 3 classes. In the calculation of accuracy we determine the probabilistic class as the class with the maximum probability among the three classes, in some cases the maximal probability can be less than 0.5, but it is always above 1/3.

The dependence of accuracy on meta-parameters of the algorithms is shown in Figs. 15 and 16. We see that for the tree-based algorithms, the optimal parameters are similar to the 2-class classification, i.e., 50 trees with depth 6 for RF and 100 trees with depth 2 for BDT provide close to optimal performance at a minimal cost in complexity (depth of the trees). The main difference for NN and LR algorithms is that more steps are needed for convergence, especially in the case of oversampling. In the following we use 600 epochs for NN and 500 iterations for LR instead of 300 epochs and 200 iterations respectively in the two-class case. For NN, the accuracy stops increasing above about 10 neurons in the hidden layer (in the following we use 11 neurons for classification: the same as in the two-class case). For oversampling, we use the oversampling factors $\sqrt{\frac{\#\ \mathrm{AGN}}{\#\ \mathrm{PSR}}}$ and $\sqrt{\frac{\#\ \mathrm{AGN}}{\#\ \mathrm{OTHER}}}$ for PSR and OTHER classes respectively (compared to the $\frac{\#\ \mathrm{AGN}}{\#\ \mathrm{PSR}}$ oversampling factor in the 2-class case). The reason for the smaller oversampling factors is to avoid overweighting the two relatively small PSR and OTHER classes.

We show an example of domains in the 3-class case in Fig. 17. A class domain is determined by the class with the largest probability. Since in the 3-class case there are two independent probabilities, which are difficult to show with a single color bar, we present only the domains represented by three different colors: brown for PSR, green for OTHER, and blue for AGN classes. The corresponding training and testing data are shown by red crosses and brown rotated squares for PSR, by green squares and yellow diamonds for OTHER, and by blue and purple triangles for AGN. The classification domains are averaged over 100 realizations of splitting the data into training and testing samples. One of these splittings is shown in the figure for illustration.

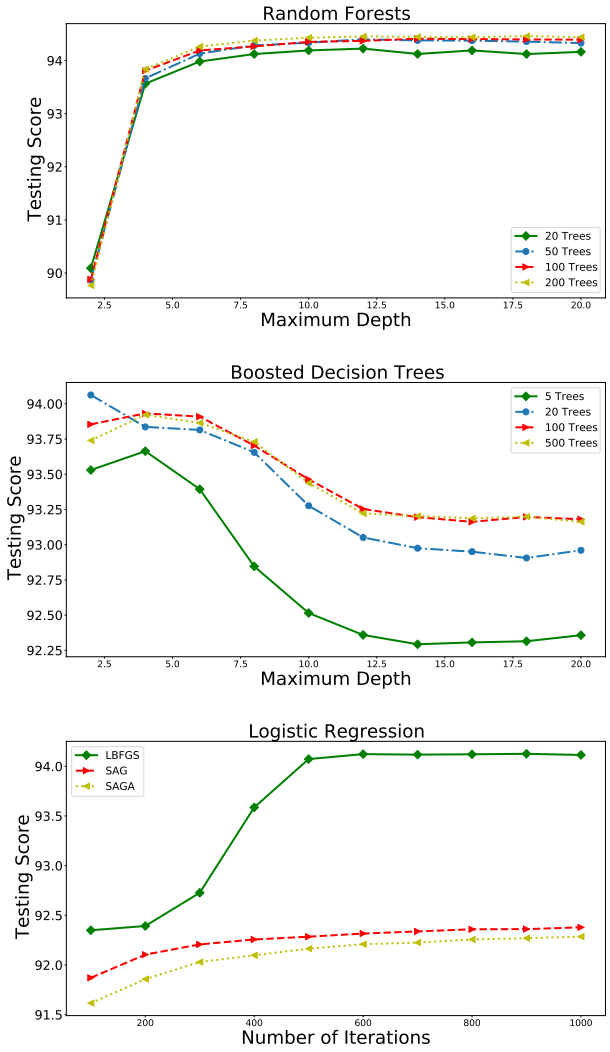

Fig. 15. Accuracy of the 3-class classification with RF, BDT and LR methods. LR does not have liblinear solver here, since liblinear does not support multi-class loss.

The accuracies of our chosen models for classification of the 3FGL sources are presented in Table 9. As in the 2-class case, the accuracies are averaged over 1000 realizations of splitting the data into training and testing samples. We notice that accuracies presented in Table 2 are calculated relative to AGN and PSR classes

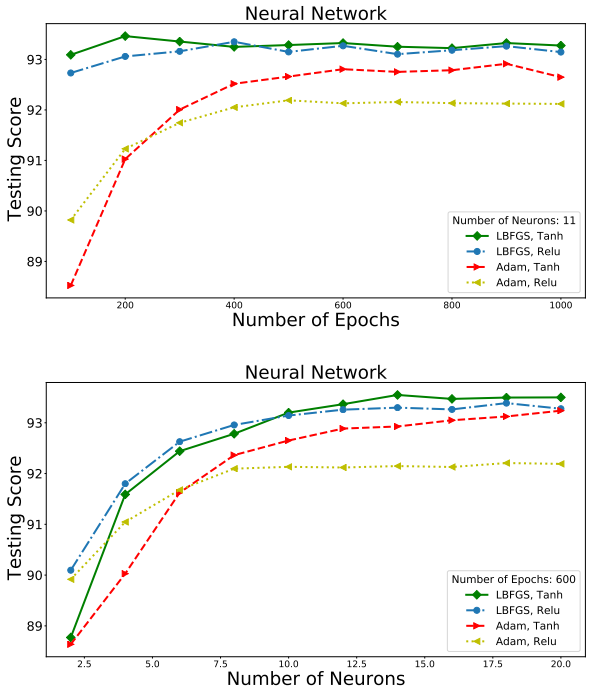

Fig. 16. Accuracy of the NN classification as a function of the number of epochs and of the number of neurons for the 3-class classification.

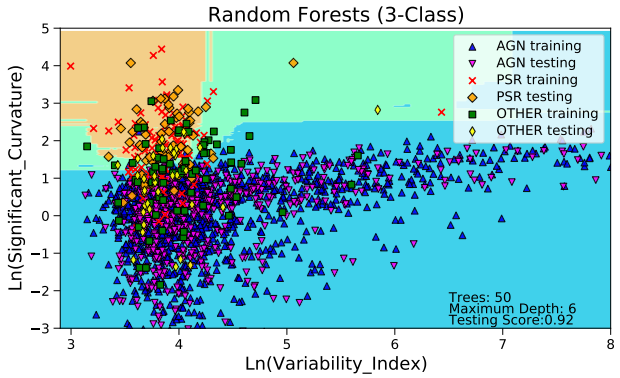

Fig. 17. Classification domains for RF in the 3-class classification. The sources in the blue, green, and brown areas are attributed to AGN, OTHER, and PSR classes respectively.

only. If we take into account that all OTHER sources are misclassified in this case, then the testing accuracy is reduced by about 5% (the fraction of OTHER sources among associated sources in 3FGL), while the accuracy of comparison with 4FGL-DR2 is reduced by about 10% (there are 37 unassociated sources in 3FGL with OTHER class associations in 4FGL-DR2, while there are in total 340 unassociated sources in 3FGL with associations in 4FGL-DR2). Thus the testing accuracy of 93-94% in Table 9 provides at least a 1-2% improvement over the accuracy in Table 2, after taking into account the misclassification of OTHER sources in the 2-class case. A similar improvement is seen for the accuracy of classification of unassociated sources in 3FGL with corresponding 4FGL-DR2 associations.

Similarly to the 2-class classification, we use the condition that all algorithms agree to determine the candidate probabilistic classes of sources. In the end of this Section we also use a stricter condition that the sum of probabilities is larger than 7 in order to determine lists of most likely PSR and OTHER source candidates. A comparison of predicted classes (using the all algorithms agreement condition) for the 3FGL unassociated sources in the 3-class classification case with the classes of the corresponding associated sources in the 4FGL-DR2 catalog are presented in Table 10. It is interesting to note that there are fewer sources with mixed classification in this case relative to the 2-class classification in Table 4. Also the number of correct predictions in the 3-class case is 263 (out of 340 sources), while in the 2-class case there are 246 correct predictions. We also show in this table the expected precision and recall for the all-algorithms-agree and $\sum_a p_a > 7$ conditions. We note that in spite of adding an additional class, the precision of the AGN and PSR classifications in 3FGL is higher in the 3-class case (Table 10) compared to the 2-class case (Table 4), but the recall is smaller for pulsars and similar or larger for AGNs in the 3-class case.

Algorithm	Parameters	Testing Accuracy	Std. Dev.	Comparison with 4FGL-DR2 Accuracy
RF	50 trees, max depth 6	93.96	0.85	85.00
RF_O		94.38	0.76	85.00
BDT	100 trees, max depth 2	93.72	0.83	83.24
BDT_O		93.83	0.80	85.29
NN	600 epochs, 11 neurons, LBFGS	93.17	1.05	83.53
NN_O		92.51	1.34	81.76
LR	500 iterations, LBFGS solver	93.93	0.88	83.24
LR_O		93.01	0.96	83.24

Table 9. Testing accuracy of the four selected algorithms for the 3-class classification of 3FGL sources and comparison with associations in the 4FGL-DR2 catalog. "_O" denotes training with oversampling.

4FGL-DR2 class	3FGL prediction				Recall	Recall	Recall
	AGN	PSR	OTHER	MIXED	(4FGL assoc)	(all agree)	($\sum_a p_a > 7$)
AGN	238	2	1	17	0.92	0.97	0.91
PSR	12	17	0	16	0.38	0.58	0.28
OTHER	6	5	8	18	0.22	0.29	–
Precision (4FGL assoc)	0.93	0.71	0.89				
Precision (all agree)	0.99	0.89	0.75				
Precision ($\sum_a p_a > 7$)	0.99	0.98	–				

Table 10. Comparison of classes predicted for unassociated sources in the 3FGL catalog using 3-class classification with associations in the 4FGL-DR2 catalog. For the definition of precision and recall labels see Table 4.

The 3-class classification of 4FGL-DR2 sources is performed analogously to the 3-class classification of the

3FGL sources. The differences between the 4FGL-DR2 and 3FGL 3-class classifications are similar to the differences in the 2-class classification of 3FGL and 4FGL-DR2 sources: we use 16 features instead of 11 features in the 3FGL case (the features are the same as in the 2-class classification of 4FGL-DR2 sources in Section 4.2 but with GLON replaced by cos(GLON)) and we have 16 neurons in the hidden layer of the NN method. Furthermore, for LR we use 1000 iterations as it gives better performance for oversampled cases. The corresponding accuracies are reported in Table 11. In comparing the accuracies with the 2-class classification in Table 6, one has to take into account that there are 346 OTHER sources among 4116 associated sources in 4FGL-DR2, which is about 8.4%. Since all OTHER sources are "misclassified" by the 2-class classification, the 3-class classification provides an improvement of about 2-4% compared to the 2-class classification.

The expected precision and recall for the classification of sources using agreement among all algorithms and the $\sum_a p_a > 7$ conditions are presented in Table 7. We note that for the 4FGL-DR2 sources the precision in the 3-class case is better than the precision in the 2-class case, while the recall is better in the 2-class case. Another comparison of the performance of the 2- and 3-class classifications is provided in Appendix D where we present the reliability diagrams. The conclusion is that the reliability diagrams for the 3-class are similar to the reliability diagrams for the 2-class case, which includes only AGN and PSR sources, with an additional advantage that in the 2-class classification one needs to account for the contribution of the OTHER sources among the unassociated ones, while in the 3-class case the contribution of the OTHER sources is included in the model.

Algorithm	Parameters	Testing Accuracy	Std. Dev.
RF	50 trees, max depth 6	92.91	0.66
RF_O		92.83	0.63
BDT	100 trees, max depth 2	92.51	0.67
BDT_O		92.27	0.67
NN	600 epochs, 16 neurons, LBFGS	91.86	0.72
NN_O		90.26	0.83
LR	1000 iterations, LBFGS solver	92.63	0.67
LR_O		92.22	0.69

Table 11. Testing accuracy of the four selected algorithms for the 3-class classification of 4FGL-DR2 sources. "_O" denotes training with oversampling.

The numbers of unassociated sources classified by all 8 methods as AGNs, pulsars, and other sources for the 3FGL and 4FGL-DR2 catalogs are presented in Table 5 in the "3-class" rows. For each algorithm the most probable class of the source is determined by the class with the largest probability. The "Mixed" column shows the number of sources with different classification results for different algorithms.

Classification of *Fermi*-LAT 4FGL sources into three classes was considered earlier by, e.g., Zhu et al. (2021), who have primarily used a two-step classification procedure, where in the first step AGNs are separated from the rest of sources and in the second step the remaining sources are split into pulsars and other sources. Zhu et al. (2021) have also tested a simultaneous classification of sources into three classes (AGN, pulsars, and other), but the results were inconsistent for the two ML algorithms used by Zhu et al. (2021) (RF and NN). In particular, the number of OTHER sources predicted by NN was zero. In our case, the predictions of various algorithms are relatively consistent with each other. For example, in the 3FGL (4FGL-DR2) catalog all 8 methods classify 69 (271) unassociated sources as OTHER. Also, 8 out of 37 unassociated 3FGL sources, which are associated to OTHER sources in 4FGL-DR2, are classified by all 8 algorithms as OTHER (see Table 10). We have also checked that the reliability diagrams for the OTHER class in the 3-class classification look reasonable (see Fig. D.2).

We also used the 3-class classification to create lists of most likely PSR and OTHER sources among the unassociated sources in the 4FGL-DR2 catalog. In Section 4.2 we determined a list of 29 PSR candidates by requiring that the unassociated sources are predicted to be pulsars by all 8 methods both in the 3FGL and in the 4FGL-DR2 catalogs. As one can see from Table 9 and also from the comparison of Tables 10 and 4, the $\sum_a p_a > 7$ condition provides a better precision than the agreement among the algorithms condition. For this reason, in this section we use the $\sum_a p_a > 7$ condition to create lists of PSR and OTHER candidates among the unassociated 4FGL-DR2 sources. In the 3FGL case, all PSR candidates satisfying the $\sum_a p_a > 7$ condition are already associated to pulsars in 4FGL-DR2, while there are no OTHER candidates satisfying the $\sum_a p_a > 7$ condition among the unassociated sources.

There are 6 unassociated 4FGL-DR2 sources with the sum of PSR-class probabilities in the 3-class classification larger than 7. The PSR candidates are shown in Table 12. All of these sources are also among the list of 29 PSR candidates determined in the 2-class classification using both 3FGL and 4FGL-DR2 features in Section 4.2. For convenience, we also add the sums of PSR-like probabilities for the 3-class classification of the 4FGL-DR2 sources in the "3FGL_4FGL-DR2_Candidates_PSR.csv" file. Three sources (4FGL J1539.4-3323, 4FGL J0933.8-6232,

Source_Name_4FGL	GLON	GLAT	$P_{\rm tot\ PSR}^{\rm 4FGL-DR2}$	$P_{\rm tot\ PSR}^{\rm 3FGL}$	3FGL cat.
4FGL J1539.4-3323	338.76	17.53	7.34	6.4	PSR
4FGL J0953.6-1509	251.94	29.6	7.28	6.84	PSR
4FGL J0933.8-6232	282.24	−7.91	7.23	6.14	PSR
4FGL J1120.0-2204	276.47	36.06	7.16	6.91	PSR
4FGL J2112.5-3043	14.9	−42.44	7.1	6.71	PSR
4FGL J1225.9+2951	185.15	83.77	7.05	5.12	MIXED

Table 12. Unassociated 4FGL-DR2 sources with the sum of PSR-class probabilities for all 8 ML methods in the 3-class classification larger than 7. All of these sources are also unassociated sources in the 3FGL catalog. $P_{\rm tot\ PSR}^{\rm 4FGL-DR2}$ ($P_{\rm tot\ PSR}^{\rm 3FGL}$) represents the sum of PSR class probabilities in the 3-class classification for the 4FGL-DR2 (3FGL) catalog. "3FGL cat." column is the probabilistic category based on the 3-class classification for the corresponding 3FGL source.

Source_Name_4FGL	GLON	GLAT	$P_{\rm tot\ OTHER}^{\rm 4FGL-DR2}$	3FGL cat.
4FGL J1800.2-2403c	5.86	−0.34	7.49	
4FGL J1417.7-6057	313.22	0.18	7.44	
4FGL J1847.7-0125	31.24	0.17	7.43	
4FGL J1838.7-0601	26.13	0.06	7.37	
4FGL J1838.4-0630c	25.65	−0.09	7.32	
4FGL J1843.7-0326	28.99	0.15	7.26	MIXED
4FGL J1556.8-5242c	328.93	0.54	7.23	
4FGL J1631.7-4826c	335.89	−0.18	7.22	OTHER
4FGL J1626.0-4917c	334.63	−0.09	7.21	OTHER
4FGL J1419.2-6029	313.54	0.56	7.21	
4FGL J1849.4-0117	31.55	−0.15	7.2	MIXED
4FGL J1109.4-6115e	290.98	−0.78	7.2	MIXED
4FGL J1850.2-0201	31	−0.67	7.18	
4FGL J1619.4-5106c	332.59	−0.62	7.17	
4FGL J1620.8-5035c	333.11	−0.42	7.15	
4FGL J1801.8-2358	6.11	−0.61	7.15	
4FGL J1511.2-5803	320.57	−0.06	7.12	
4FGL J1800.9-2407	5.88	−0.51	7.12	
4FGL J1312.6-6231c	305.37	0.24	7.12	OTHER
4FGL J1357.3-6123	310.74	0.48	7.11	
4FGL J1353.5-6128	310.27	0.52	7.08	
4FGL J1756.6-2352	5.62	0.47	7.06	
4FGL J1855.8+0150	35.07	−0.14	7.05	
4FGL J1650.9-4420c	341.15	0.04	7.03	
4FGL J1742.0-3020	358.4	−0.07	7.01	
4FGL J1904.7+0615	40.01	−0.09	7.01	
4FGL J1618.0-5119	332.28	−0.63	7.01	
4FGL J1620.8-4958c	333.55	0.03	7	
4FGL J1321.1-6239	306.34	0.02	7	
4FGL J1646.5-4406	340.82	0.8	7	

Table 13. Unassociated 4FGL-DR2 sources with the sum of OTHER-class probabilities for all 8 ML methods in the 3-class classification larger than 7. For the description of columns, see Table 12. Sources with missing values in the 3FGL columns are not detected in the 3FGL. We also save these OTHER-class candidates in the supplementary online materials as "4FGL-DR2_Candidates_OTHER_3classes.csv" (SOM 2021).

and 4FGL J2112.5-3043) have also possible associations in the Parkes survey (Table 8).

There are 30 unassociated 4FGL-DR2 sources with the sum of OTHER-class probabilities in the 3-class classification larger than 7. The OTHER-class candidates are shown in Table 13. Out of the 30 sources only two sources have no flags in 4FGL, and only eight sources have an association with the previous FGL catalogs (column name ASSOC_FGL). The following sources had additional associations in the Simbad database ordered by decreasing sum of OTHER-class probabilities:

1. 4FGL J1800.2-2403c: This source has the largest sum of OTHER-class probabilities but no entry in Simbad. However, the source 1FGL J1800.5-2359c (Abdo et al. 2010a) is associated to two sources in 4FGL-DR2: 4FGL J1800.7-2355, which has OTHER association in the region of the SNR W28 (Ritchey 2020), and 4FGL J1800.2-2403c.
2. 4FGL J1847.7-0125: Within 1 arcminute of the candidate Young Stellar Object (YSO) SSTOERC G031.2256+00.1711 (Saral et al. 2017).
3. 4FGL J1843.7-0326: Associated with 3FGL J1843.7-0322 and found near the HESS source HESS J1843-033, next to the SNR G28.6-0.1 (H. E. S. S. Collaboration et al. 2018). This source is among the 120 unassociated sources according to significance (>10) in the list of Saz Parkinson et al. (2016) where the RF and LR methods predicted it to be a young pulsar based on the 2-class classification. In our 3FGL 3-class catalog, this source has a 'MIXED' classification.
4. 4FGL J1556.8-5242c: Within 1 arcminute of the candidate YSO 2MASS J15564953-5241450 (Robitaille et al. 2008).
5. 4FGL J1631.7-4826c (3FGL J1632.4-4820 (Acero et al. 2015)): Within 30 arcseconds of the Dark Cloud (Nebula) SDC G335.894-0.184 (Peretto et al. 2016).
6. 4FGL J1626.0-4917c (3FGL J1626.2-4911 (Acero et al. 2015), 3FHL J1626.3-4915 (Ajello et al. 2017)): Associated with HESS J1626-490. It is also one of the 27 sources shortlisted by Hui et al. (2020), who used ML methods to select pulsar candidates from the 3FHL catalog. It is also classified as OTHER based on the 3FGL values.
7. 4FGL J1849.4-0117 (3FGL J1849.5-0124c (Acero et al. 2015)): It is in the region of Galactic mini starburst W43 studied by Yang & Wang (2020). Also within 1 arcminute of the candidate YSO SSTOERC G031.5367-00.1555. Has a 'MIXED' classification based on the 3FGL values.
8. 4FGL J1109.4-6115e (3FGL J1111.9-6038 (Acero et al. 2015)): It is associated with the extended galactic source FGES J1109.4-6115 (Ackermann et al. 2017), near the speculated SFR 4FGL J1115.1-6118, in there region of Young Massive Stellar Cluster NGC 3603 (Saha et al. 2020). 'MIXED' classification with the 3FGL values.
9. 4FGL J1850.2-0201: Also in the region of the starburst W43 (Yang & Wang 2020).
10. 4FGL J1801.8-2358: Associated with HESS J1800-240A and 2FHL J1801.7-2358. Located south of the SNR W28 (Ritchey 2020).
11. 4FGL J1855.8+0150: In the region of SNR W44 (Peron et al. 2020).
12. 4FGL J1742.0-3020: Within 1 arcminute of the Molecular cloud [MML2017] 777 (Miville-Deschênes et al. 2017).

13. 4FGL J1904.7+0615: Within 1 arcminute of the Bubble [SPK2012] MWP1G040012-001102 (Simpson et al. 2012).

6. Application of probabilistic catalogs for population studies

6.1. Number of sources as a function of flux

In this section we show how probabilistic catalogs can be used for population studies. One of the most important questions in gamma-ray astronomy is the contribution of point sources, e.g., AGNs, to the extragalactic gamma-ray flux (e.g., Abdo et al. 2010b; Malyshev & Hogg 2011; Ackermann et al. 2016; Zechlin et al. 2016b,a; Lisanti et al. 2016; Di Mauro et al. 2018): if most of the extragalactic emission is explained by point sources, then one can put stringent constraints, e.g., on dark matter annihilation or decay into gamma rays (Ajello et al. 2015; Di Mauro & Donato 2015; Ackermann et al. 2015; Fornasa & Sánchez-Conde 2015; Liu et al. 2017) or on the evaporation of primordial black holes (Carr et al. 2010). In particular, it is important to understand the contribution to the population of AGNs from the unassociated sources. A probabilistic catalog provides an answer to the question: how many sources among the unassociated ones are expected to belong to different classes, such as pulsars or AGNs. One can calculate the total expected number of AGNs or pulsars among the unassociated sources, or calculate the contribution as a function of one or more parameters. In this subsection we determine the numbers of AGNs, pulsars, and, in case of 3-class classification, OTHER sources as a function of their flux. In the following subsection we also discuss the distributions of sources as functions of Galactic latitude and longitude.

In Fig. 18 we show the cumulative number of AGNs and pulsars with a flux above 1 GeV larger than the value on the x-axis. Solid blue lines show the actual counts of sources (AGNs or pulsars) in the 3FGL and 4FGL-DR2 catalogs. As a consistency check of the method, we calculate the AGN- and PSR-like probabilities for associated sources. The sum of probabilities (uncorrected for sources other than AGNs and pulsars) for LR algorithm are shown by dotted purple lines. In order to correct the expected number of AGNs among associated sources for AGN-like probabilities in "other" sources, we subtract the corresponding AGN-like probabilities in each flux band:

$$N_{\rm AGN}^{\rm ass} = \sum_{i\in{\rm ass}} p^i_{\rm AGN} \; - \sum_{i\in{\rm ass\,other}} p^i_{\rm AGN}. \qquad (8)$$

The corrected sums of probabilities for LR method are shown by dashed purple lines. The green bands show the envelope of the sums of corrected probabilities for the eight methods used in this paper. We see that the counts of associated sources, AGNs and pulsars, are consistent with the expected number of associated sources calculated from the class probabilities of associated sources. This conclusion is not very surprising since we used associated sources for training the ML algorithms. We note that the correction for "other" sources is important for consistency of the sum of probabilities and the number of associated sources. We have also compared the sums of probabilities for the 3FGL associated sources in Saz Parkinson et al. (2016). The sum of probabilities for associated sources in the LR case uncorrected for "other" sources are shown by dotted black line, while the sums corrected for "other" sources are shown by black dash-dotted lines. The gray band is the envelope of the two methods (LR and RF) used by Saz Parkinson et al. (2016). We see that the sum of probabilities for AGNs and pulsars overpredicts the source counts in 3FGL, while correction for "other" sources makes the prediction for associated sources consistent with the source counts.

The predictions for the number of AGNs and pulsars among the unassociated sources corrected for "other" sources added to the 3FGL and 4FGL-DR2 source counts are shown by solid orange lines for the LR model. The orange bands show the corresponding envelopes for the eight ML methods. We assume that the fractional contribution of other sources is the same for associated and unassociated sources in the different flux bands. Thus, the correction for the presence of other sources is calculated similarly to the associated sources in Eq. (8), but we adjust for the fact that there are fewer unassociated than associated sources, i.e., the correction is assumed to be proportionally smaller. In particular, the number of AGNs among unassociated sources in a flux band ΔF is estimated as

$$N_{\rm AGN}^{\rm unass} = \sum_{i\in{\rm unass}} p^i_{\rm AGN} \; - \sum_{i\in{\rm ass\,other}} p^i_{\rm AGN} \cdot \frac{N_{\rm unass}}{N_{\rm ass}} \qquad (9)$$

where all probabilities and the number of sources are computed for sources with flux inside ΔF. The first term is the sum of AGN-like probabilities among the unassociated sources, while the second term is the sum of AGN-like probabilities among associated "other" sources rescaled by the total number of unassociated and associated sources in this flux band. The expected number of pulsars among the unassociated sources is calculated analogously. The corresponding sums of associated source counts plus the expected number of sources calculated with the LR method of Saz Parkinson et al. (2016) and corrected for other sources are shown by dashed grey lines.

The expected numbers of pulsars and AGNs among the unassociated sources are summarized in Table 14 for

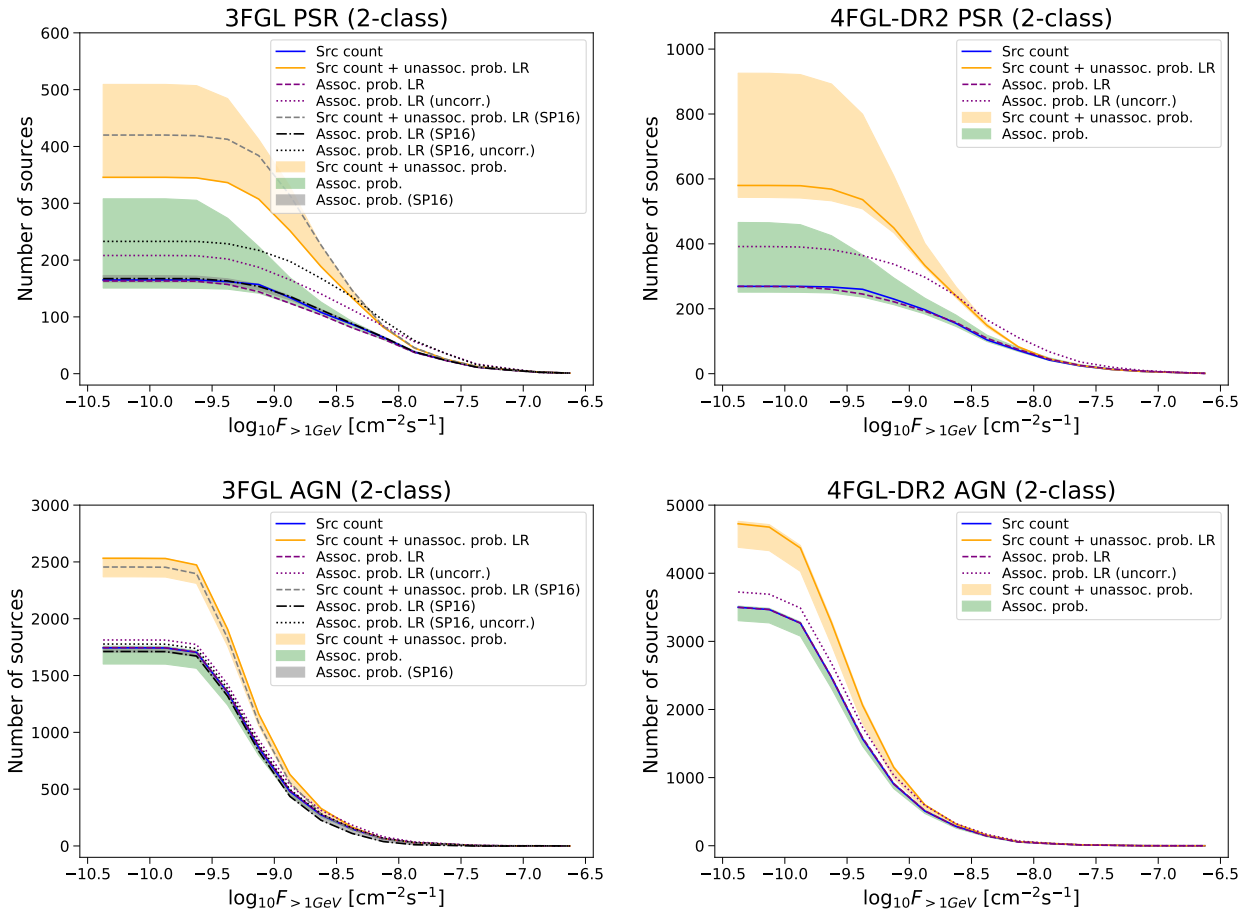

Fig. 18. Cumulative number of sources as a function of their flux. Blue solid line – associated 3FGL and 4FGL-DR2 sources. Green band – envelope of sums of class probabilities for associated sources for the eight ML methods corrected for the presence of OTHER sources. Orange solid line (band) – sum of class probabilities for the LR model (the envelope of the eight ML methods), corrected for the presence of OTHER sources, added to the source count of associated sources. Purple dashed (dotted) line – sum of class probabilities for associated sources for the LR method without oversampling corrected (uncorrected) for the presence of OTHER sources. Gray dash-dotted (dotted) line – sums of class probabilities from Saz Parkinson et al. (2016) using LR corrected (uncorrected) for the presence of OTHER sources. Gray band – envelope of sums of class probabilities for associated sources for LR and RF methods from Saz Parkinson et al. (2016) corrected for the presence of OTHER sources. For details see Section 6.1.

the 3FGL catalog and in Table 15 for the 4FGL-DR2 catalog. The row "2-class" shows the expectations for the 2-class classification uncorrected for the presence of other sources. The row "2-class corr total" shows the expectations corrected for other sources using Eq. (9) without any binning, while "2-class corr F-bins" shows the correction with the flux bins as in Fig. 18. We see that corrections using total source counts and source counts binned in flux are similar and relatively small. In the following section we also calculate the correction using latitude and longitude bins.

The numbers of PSRs and OTHER sources in the 3-class classification as a function of flux are shown in Fig. 19. As in the 2-class case, solid blue line shows the counts of associated sources in the corresponding classes. Red and blue bands show the envelops of the expected number of associated sources and the number of associated sources plus the expected numbers of sources among unassociated ones respectively. In the PSR plots, we also replot the envelops of expected numbers of associated PSR (shown by the green band) and associated source counts plus predictions for unassociated sources by orange bands corrected for the presence of OTHER sources among unassociated ones. We notice that the bands for the 3-class case are narrower in the PSR plots than the bands for the 2-class case. In part this is due to less oversampling in the 3-class case. We also note that the red band lies almost entirely below the orange one. This is due to a possible underestimation of the contribution of OTHER sources to the PSR class among unassociated sources in the 2-class case. In particular, in the bottom panels of Fig. 19 we show the estimated total number of OTHER sources in the 2-class case by the orange line. Since we do not have probabilities for the OTHER sources in this case, we estimate the number of OTHER sources among unassociated ones simply by

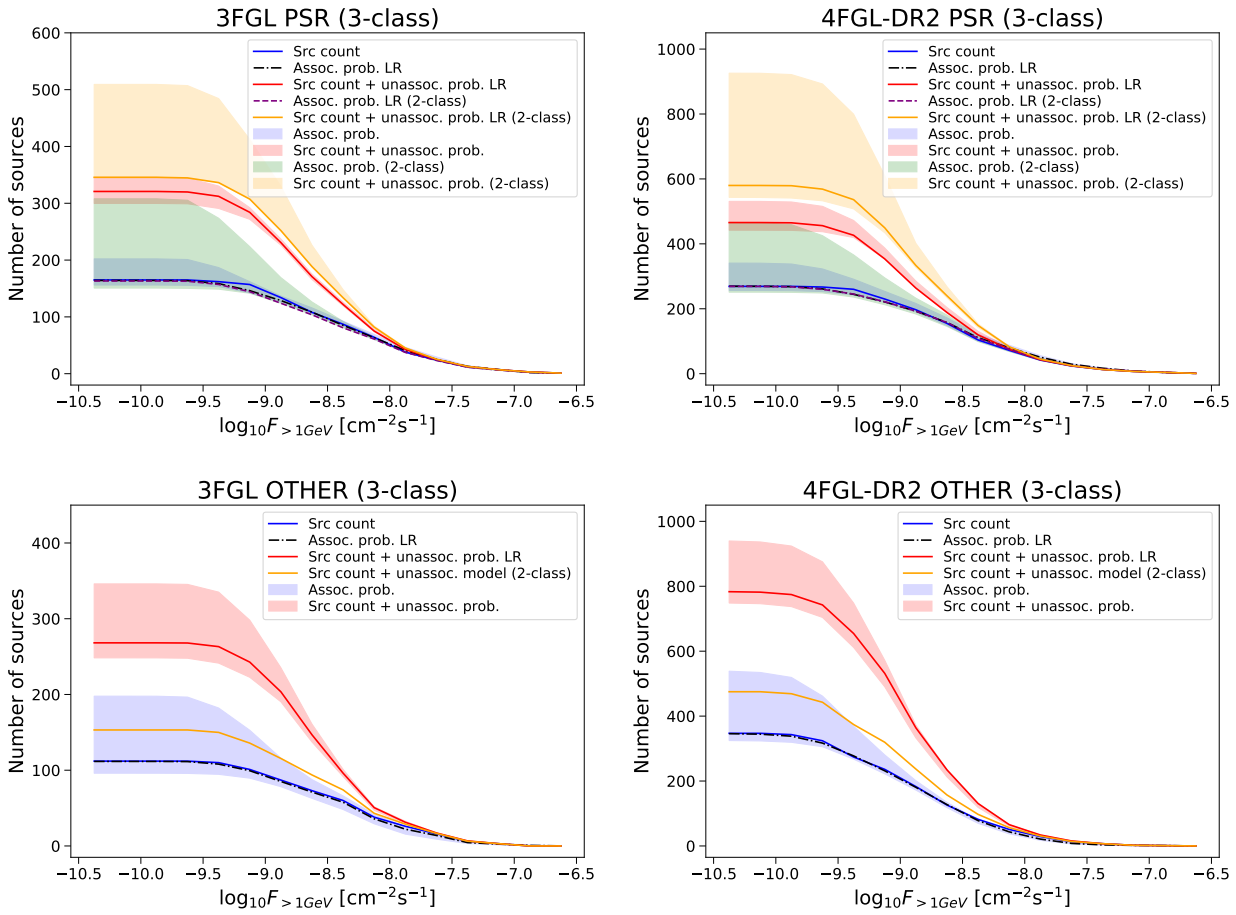

Fig. 19. Cumulative number of sources as a function of their flux. Blue solid line – associated 3FGL and 4FGL-DR2 sources. Blue band – envelope of sums of class probabilities for associated sources for the eight ML methods. Purple dash-dotted line – sum of class probabilities for associated sources for the LR method without oversampling. Red solid line (band) – sum of class probabilities for the LR model (the envelope of the eight ML methods) added to the source count of associated sources. Purple dashed line – sum of class probabilities for associated sources for the LR method without oversampling in the 2-class classification corrected for the presence of OTHER sources. Orange solid line (band) – sum of class probabilities for the LR model (the envelope of the eight ML methods), corrected for the presence of OTHER sources, added to the source count of associated sources (the band and the line are the same as in Fig. 18). Green band in the PSR plots – envelope of sums of class probabilities for associated sources for the eight ML methods in the 2-class classification corrected for the presence of OTHER sources (the same as in Fig. 18). Orange solid line in the OTHER plots – OTHER source counts plus an estimate of the number of OTHER sources among unassociated ones using Eq. (10). For details see Section 6.1.

rescaling the number of associated OTHER sources in each energy band as

$$N^{\rm unass}_{\rm OTHER} = N^{\rm ass}_{\rm OTHER} \frac{N^{\rm unass}}{N^{\rm ass}} \tag{10}$$

We notice that this estimate agrees with the correction of the estimated numbers of AGNs and pulsars in the 2-class case due to presence of other sources, since

$$\begin{aligned} N^{\rm ass}_{\rm OTHER} &= \sum_{i\in {\rm ass\,other}} p^i_{\rm AGN} + \sum_{i\in {\rm ass\,other}} p^i_{\rm PSR} \\ &= \sum_{i\in {\rm ass\,other}} (p^i_{\rm AGN} + p^i_{\rm PSR}) \end{aligned} \tag{11}$$

provided that $p^i_{\rm AGN} + p^i_{\rm PSR} = 1$ for each source in the two-class case. The estimated total number of OTHER sources in the 2-class case (orange line) is significantly below the red band derived from the sum of probabilities of the OTHER class in the 3-class case. The model in Eq. (10) depends on binning: as we show in the following section, for latitude binning this 2-class estimate of the number of OTHER sources among unassociated ones agrees with the 3-class estimate. The reason is that for flux bins, the ratio $\frac{N^{\rm unass}}{N^{\rm ass}}$ is dominated by AGNs at high latitudes and it is always smaller than 1, while at low latitudes, where most of OTHER sources are, the ratio $\frac{N^{\rm unass}}{N^{\rm ass}}$ is about 1.

We note that the probabilistic classification mostly affects sources with small fluxes. In Fig. 20 we plot the ratio of the expected number of sources of a certain class among unassociated sources computed according to Eq. (9) with the LR algorithm (without oversampling) to the number of associated sources in this class. The

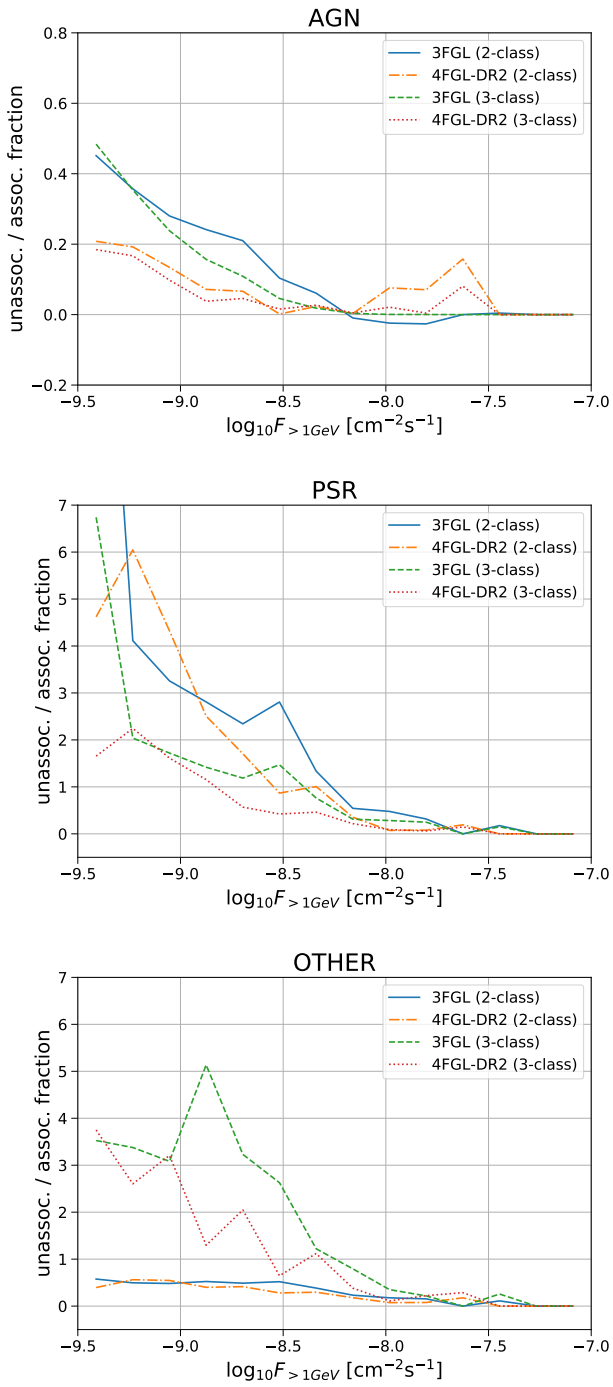

Fig. 20. Ratio of the number of AGNs, pulsars, and other sources among unassociated sources estimated with LR without oversampling to the counts of associated sources respectively. AGN and pulsar estimates are corrected for the presence of other sources using Eq. (9), while the number of other sources among unassociated ones in the 2-class case is estimated using Eq. (10).

ratio generally increases as the flux decreases. Negative values (e.g., at high fluxes for AGNs) are due to subtraction of probabilities for the "other" associated sources. For OTHER sources in the 2-class case we use estimates in Eq. (10). As discussed above, we see that these estimates are significantly below the estimated numbers of OTHER sources among unassociated ones in the 3-class classification (for the LR method without oversampling in this case).

6.2. Latitude and longitude profiles

In this section we show the Galactic latitude and longitude profiles of the distributions of associated and unassociated sources. In Figs. 21 and 22 we present the source counts as a function of abs(sin(GLAT)) for 2-class and 3-class classifications respectively. We use 20 bins, i.e., each bin corresponds to a solid angle of $4\pi/20$. Solid blue lines show counts of associated sources in 3FGL and 4FGL-DR2 catalogs. The total counts of unassociated sources are shown by red dash-dotted lines. Green bands show the envelopes of sums of probabilities for classification of unassociated sources into AGN, PSR, and, in case of 3 classes, OTHER classes for the eight ML methods corrected in the case of 2 classes for the presence of OTHER sources. Black dotted line in plots of the OTHER class show the model for the number of OTHER sources among the unassociated ones in Eq. (10) for latitude bins. We see that in the latitude profile, the 2-class model for the contribution of OTHER sources to the unassociated ones is generally consistent with the estimate of the number of OTHER sources in the 3-class model.

The classifications of unassociated 3FGL sources by Saz Parkinson et al. (2016) are shown by gray dashed (RF) and dotted (LR) lines. The numbers of unassociated sources classified as AGN, PSR, or OTHER grow towards the Galactic plane (GP). Within $\approx 3°$ from the GP the expected number of PSRs is about the same as the number of AGNs among unassociated sources, while at high latitudes, most of the unassociated sources are classified as AGNs. It is interesting to note, that according to Table 1, GLAT is one of the least important features for the RF and BDT algorithms. It can be a posteriori explained by the fact that the density of AGNs is so high that even in the GP the expected number of AGNs is comparable to the expected number of PSRs.

Orange shaded areas show the sum of the source counts and the expected number of sources for the eight methods (both with and without oversampling). The average among the eight methods added to the counts of associated sources is shown by solid orange line (for AGNs we also show the mean of these points by dotted green line). We find that the number of associated AGNs is decreasing towards the GP, the expected number of AGNs among unassociated sources is increasing towards the GP, so that the sum of the two is relatively uniform as a function of Galactic latitude.

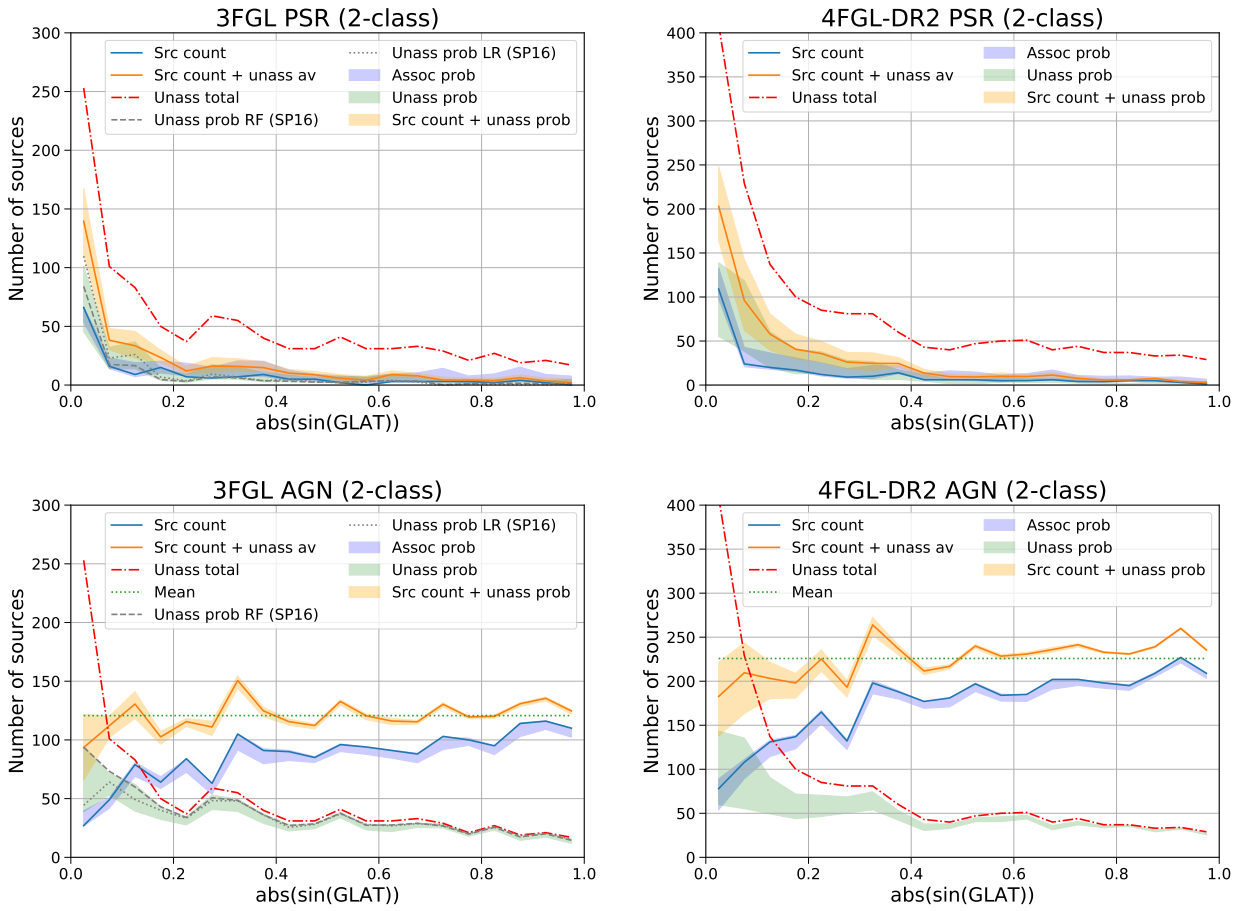

Fig. 21. Latitude profiles of source counts in case of 2-class classification. Blue solid line – associated 3FGL and 4FGL-DR2 sources. Red dash-dotted line – counts of all unassociated sources. Blue band - envelope of sums of class probabilities for associated sources for the eight ML methods with and without oversampling corrected for the presence of OTHER sources. Green band – envelope of sums of class probabilities for unassociated sources for the eight ML methods corrected for the presence of OTHER sources. Orange solid line (band) – average (envelope) of sums of class probabilities for the eight ML methods added to the source count of associated sources. Green dotted line on the AGN plots – mean of the orange solid line. Gray dashed (dotted) line – RF (LR) sums of class probabilities from Saz Parkinson et al. (2016). For details see Section 6.2.

In Figs. 23 and 24 we show plots analogous to Figs. 21 and 22 for Galactic longitudes (we use 30 bins). We note that there is a significant increase in the number of unassociated sources in the 4FGL-DR2 catalog for $|\ell| \lesssim 50°$. In the 2-class classification about half of the unassociated sources at these longitudes are attributed to pulsars and half to AGNs. In the 3-class classification at least one third of the unassociated sources in this range of longitudes is attributed to the OTHER class. This comes mostly at the expense of reducing the predicted number of pulsars. It is interesting to note that the 2-class model for the OTHER sources calculated using Eq. (10) for longitude bins (black dotted line in the OTHER plots) is significantly below the bands of 3-class expectations, whereas for the latitude bins the 2-class model is consistent with the bands. The reason is that the ratio of unassociated to associated sources, which we use to estimate the number of OTHER sources among the unassociated ones, depends on binning. In particular, this ratio is about 1 for low latitudes, which leads to an estimate that at low latitudes the number of OTHER sources among unassociated ones is similar to the number of associated OTHER sources. But for longitude and flux binning the ratio of unassociated to associated sources is smaller than 1, which leads to a prediction (in the 2-class case) that the number of OTHER sources among unassociated ones is smaller than the number of associated OTHER sources.

We summarize the expected numbers of sources in the 2-class and 3-class models in Table 14 for the 3FGL and in Table 15 for the 4FGL-DR2 catalogs. In the 2-class case we also show the correction for the presence of OTHER sources among unassociated ones. We make the correction using the total numbers of unassociated and associated sources, as well as binning in flux from Fig. 18 (F-bins), in latitudes from Fig. 21 (Lat-bins), and in longitudes from Fig. 23 (Lon-bins). We see that predictions for the numbers of sources using latitude binning

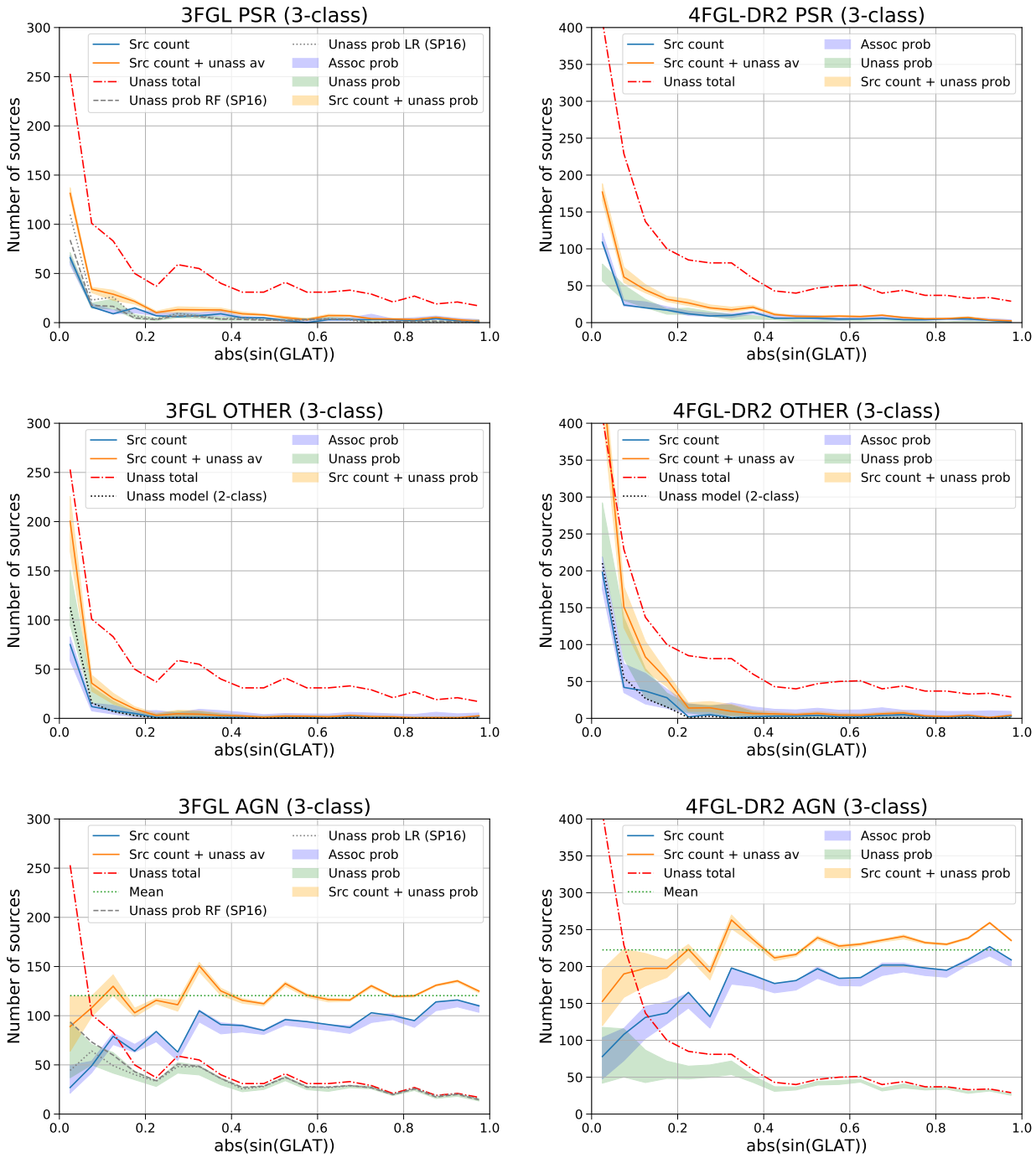

Fig. 22. Latitude profiles of source counts in case of 3-class classification. For the definition of labels see Fig. 21. Black dashed line in plots of the OTHER class shows the model for the number of OTHER sources among the unassociated ones in Eq. (10) for latitude bins.

in the 2-class case is closest to the 3-class case. Even better agreement would likely be achieved if one uses a simultaneous binning in flux, latitudes, longitudes and, possibly, other variables. If, in addition, one allows bins of variable size and non-rectangular boundaries in multi-dimensional space, then one will likely end up with one of the ML algorithms for classification. Indeed, ML algorithms are designed to provide an optimal binning, which maximizes the separation of classes and make predictions for data with unknown labels (e.g., unassociated sources) based on counts of samples in bins with known labels (e.g., associated sources). Thus the 3-class prediction for classification of unassociated sources is likely more accurate than the 2-class classification with correction for OTHER sources based on ad hoc binning of one of the variables. The 2-class classification may still be useful in situations, when high recall is necessary for either pulsars or AGNs: since OTHER sources are mixed with pulsars

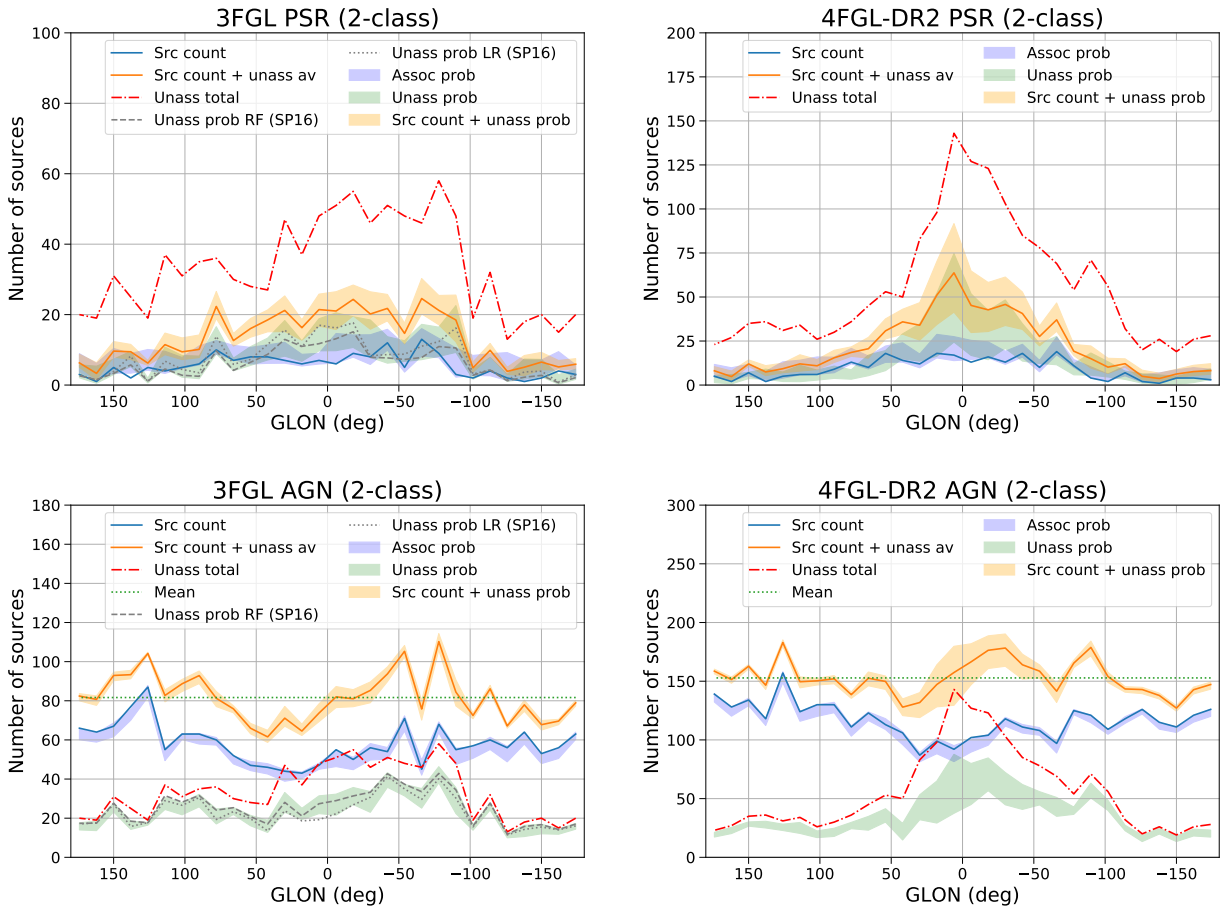

Fig. 23. Longitude profiles of source counts in case of 2-class classification. For the definition of labels see Fig. 21.

and AGNs (cf. Fig. 17), some of pulsars and AGNs can be mistakenly classified as OTHER in the 3-class case.

Classification	AGN	PSR	OTHER
2-class	$740.3^{+75.8}_{-97.6}$	$269.7^{+97.6}_{-75.8}$	–
2-class corr total	$711.6^{+70.6}_{-90.9}$	$242.0^{+90.9}_{-70.6}$	56.4
2-class corr F-bins	$717.3^{+70.8}_{-92.7}$	$251.5^{+92.7}_{-70.8}$	41.1
2-class corr Lat-bins	$670.2^{+64.9}_{-80.9}$	$196.5^{+80.9}_{-64.9}$	143.3
2-class corr Lon-bins	$705.7^{+68.3}_{-89.8}$	$234.8^{+89.8}_{-68.3}$	69.4
3-class	$664.0^{+74.9}_{-65.2}$	$158.3^{+22.3}_{-23.6}$	$187.6^{+46.4}_{-51.3}$

Table 14. Expected counts of sources among unassociated 3FGL sources. 2-class corr total (F-bins, Lat-bins, Lon-bins) shows the 2-class expectations corrected for the presence of OTHER sources using the ratio of unassociated and associated OTHER sources for the total counts (using flux, latitude, longitude bins respectively) in Eq. (9). The number of OTHER sources in the 2-class case is estimated in Eq. (10).

Classification	AGN	PSR	OTHER
2-class	$1184.4^{+174.3}_{-244.9}$	$482.6^{+244.9}_{-174.3}$	–
2-class corr total	$1104.2^{+160.0}_{-221.5}$	$422.4^{+221.5}_{-160.0}$	140.4
2-class corr F-bins	$1103.4^{+159.3}_{-223.3}$	$433.5^{+223.3}_{-159.3}$	128.1
2-class corr Lat-bins	$1013.2^{+142.6}_{-191.0}$	$334.6^{+191.0}_{-142.6}$	319.2
2-class corr Lon-bins	$1081.0^{+155.3}_{-211.6}$	$390.0^{+211.6}_{-155.3}$	196.0
3-class	$948.0^{+122.7}_{-135.8}$	$214.9^{+47.0}_{-41.8}$	$504.1^{+88.9}_{-102.6}$

Table 15. Expected counts of sources among unassociated 4FGL-DR2 sources. For the description of rows, see Table 14.

7. Conclusions

In this paper we determine the probabilities of classification of sources in the 3FGL and 4FGL-DR2 *Fermi*-LAT catalogs into two (AGNs and pulsars) and three (AGNs, pulsars, and other sources) classes. The probabilities are calculated with 8 different ML methods: RF, BDT, LR and NN – each algorithm with and without oversampling during training. The algorithms were trained and tested with associated sources. We optimized meta-parameters of the algorithms, such as the depth of the trees, the number of trees, the number of neurons, etc. to avoid any overfitting of data while providing a good accuracy of classification. The testing accuracies of classification of associated sources in the 3FGL catalog for the four algorithms in the 2-class case without (with) oversampling are about 97% (between 93% and 97%).

We have also checked the accuracy of classification by selecting unassociated sources in 3FGL, which have associations in 4FGL-DR2. If we take the 4FGL-DR2 associations as the true classes, then the accuracies of classi-

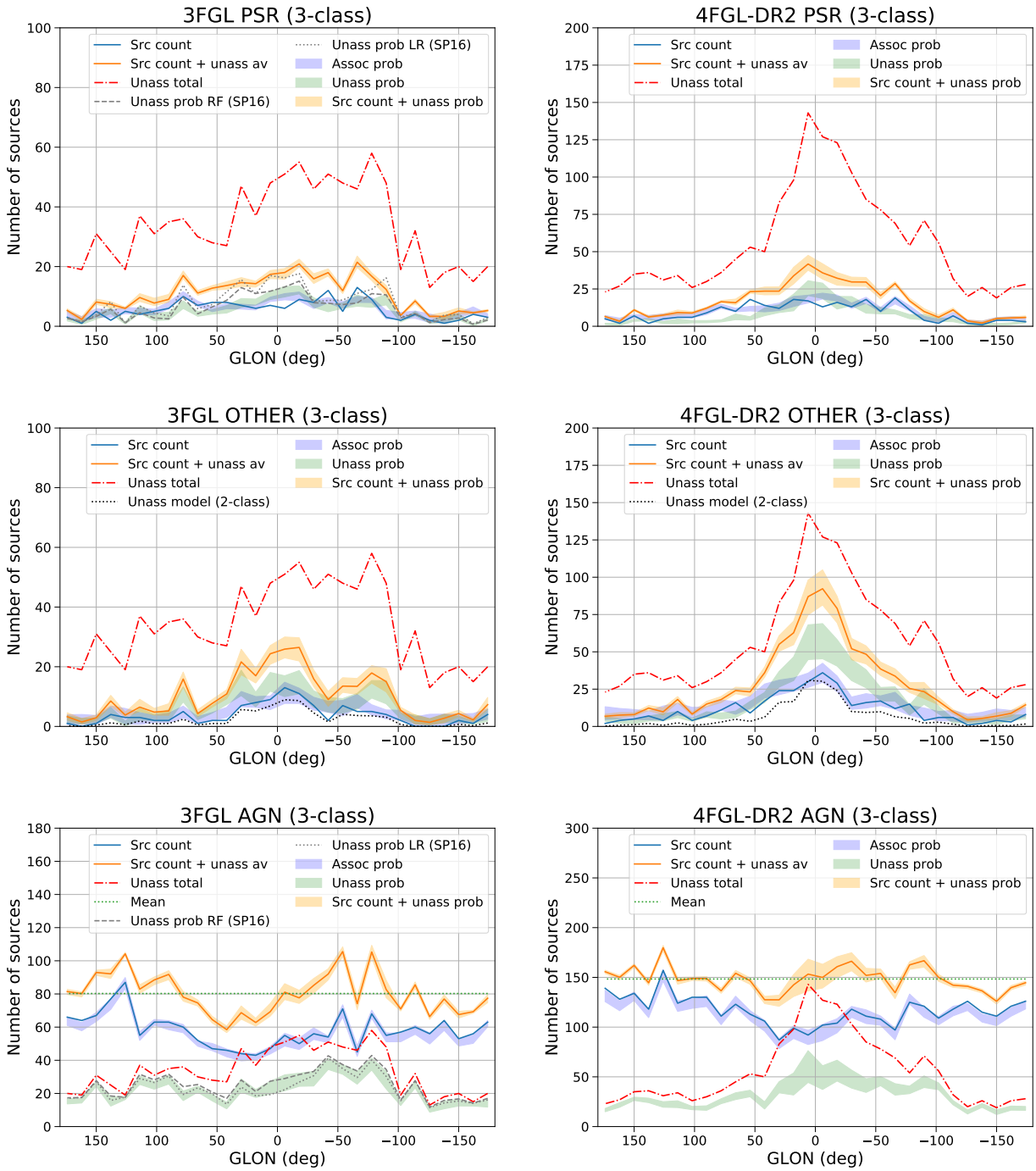

Fig. 24. Longitude profiles of source counts in case of 3-class classification. For the definition of labels see Figs. 21 and 22.

fication in this subset of sources without (with) oversampling are between 90% and 91% (85% and 92%). Most of misclassified sources in this comparison have spectral parameters in 3FGL which are typical of the other class (Fig. 14), i.e., the misclassification can be due to problems with reconstructing the spectrum of the sources. In the 3-class classification, the testing accuracies for the 3FGL catalog are between 92% and 94% (both with and without oversampling), while comparison with 4FGL-DR2 give accuracy between 82% and 85%. For the 4FGL-DR2 classifications, the testing accuracy is between 90% and 93% (both with and without oversampling). If one takes into account that all other sources are misclassified in the 2-class case, then the 3-class case provides an improvement in accuracy of 1% to 5%.

We have created four catalogs with probabilistic classifications of sources: based on 3FGL and on 4FGL-DR2 with 2- and 3-class classifications. For each source and for each class we report class probabilities for each of the eight ML methods (with and without oversampling).

We also provide individual standard deviations for all classification probabilities by using sample average over selection of training and testing datasets. We report the classification probabilities not only for the unassociated sources, but also for the associated ones, which can be used to find outliers. The full probabilistic catalogs for 2- and 3-class classifications of the 3FGL and 4FGL-DR2 catalogs are available online (SOM 2021). An advantage of such probabilistic classification is that a threshold on probability for selecting, e.g., pulsar candidates, can be chosen by the user based on their needs. For example, in a search for new pulsars, one can select a low threshold in order to avoid missing possible pulsars. In a derivation of an average property of the class, e.g., spectral index or cutoff energy, one can select a high threshold in order to avoid contamination from the other classes.

We discuss two examples of applications of probabilistic catalogs: determination of most likely classes of unassociated sources (which can be used for searches of new class members, such as AGNs or pulsars) and population studies using class probabilities including both associated and unassociated sources. For the determination of the candidate classes of unassociated sources we use two conditions: agreement among algorithms, i.e., each algorithm predicts the same class as the most likely for a source, and that the sum of probabilities for the 8 classification methods is above 7. In order to evaluate the performance of these classification conditions, we estimate the precision and recall using associated sources and test the estimations using unassociated 3FGL sources, which have associations in the 4FGL-DR2 catalog. ~~We find that the expected precision for the AGNs (pulsars) is above about 97% (80%) for 2- and 3-class classification of both 3FGL~~ Both precision and recall for classification of unassociated 3FGL sources with 4FGL-DR2 ~~catalogs. The precision for the OTHER sources in the 3-class classification is between 75% and 85%. The precision~~ associations are smaller than the estimates based on testing samples. For instance, the precision for ANGs (pulsars) estimated from the comparison of ~~unassociated~~ 3FGL ~~sources with~~ and 4FGL-DR2 ~~associations~~ catalogs is about 93% (~~70%) for AGNs (pulsars), which is significantly smaller than the precision~~ 68%) compared to the precision of 97% (78%) determined from the ~~probabilistic classes of the associated~~ testing samples for the 2-class classification of 3FGL sources (Table 4). Similar reduction of precision is observed for the 3-class classification of the 3FGL sources. We notice that many misclassified sources in the 3FGL versus 4FGL-DR2 comparison (Fig. 14) are outside of the expected class domains, if we use 4FGL-DR2 features instead of the 3FGL features. It shows that in many cases such misclassification is due to errors or uncertainties in the input data rather than issues in the classification methods. Such errors in the input data provide an irreducible uncertainty in the analysis. As a result, we find that precision and recall determined from the comparison of the 3FGL and 4FGL-DR2 catalogs provide a more realistic estimate of the true precision and recall than the values based on testing samples. Although a comparison of the 4FGL-DR2 catalog with a newer *Fermi*-LAT catalog is not possible at the moment, a similar reduction of precision and recall compared to the values determined from testing samples in the 4FGL-DR2 case can be expected.

We use the all-algorithms-agree method for the general assignment of the candidate classes to the sources (for convenience we add these classes in the catalogs) and also to create a list of 29 pulsar candidates based on unassociated sources classified as pulsars by all algorithms both in 3FGL and 4FGL-DR2 catalogs using 2-class classification. We find that the expected precision for the "sum of probabilities above 7" condition is larger than the "all-methods-agreement" condition (albeit at the expense of generally smaller recall), thus the "sum of probabilities above 7" condition is stricter. In particular, all PSR candidate sources in 3FGL satisfying this condition are associated to pulsars in 4FGL-DR2, while there are no OTHER sources in 3FGL satisfying this condition. In the 3-class classification based on the 4FGL-DR2 catalog there are 6 pulsar candidates and 30 OTHER sources candidates among unassociated sources satisfying the "sum of probabilities above 7" condition. We report these sources in Tables 12 and 13 respectively (digital versions of the tables are available in the supplementary online materials (SOM 2021)). We discuss possible associations of the pulsar candidates with pulsars from Parkes survey (Camilo et al. 2015) and OTHER sources candidates with sources in Simbad database.

As the second example of the application of the probabilistic catalogs, we perform population studies using class probabilities both for associated and unassociated sources. In particular we derive the expected number of sources in the catalog as a function of their flux. As a consistency check, we compare the counts of associated sources to the sums of probabilities for the associated sources. We find that correcting for the contribution of other sources in the 2-class case plays an important role for the estimation of the expected number of sources in a particular class. We find the total expected number of AGNs and pulsars in the 3FGL and 4FGL catalogs by adding the class probabilities for the unassociated sources in the 2- and 3-class cases to the source counts of associated sources and correcting in the 2-class case for the contribution of other classes in the unassociated sources. In particular, we find that the total expected

number of pulsars is about two times larger than the number of associated pulsars.

We also plot the counts of associated sources and the expected numbers of AGNs, pulsars, and other sources among unassociated sources as a function of Galactic latitude and longitude. We find that the number of associated AGNs is decreasing towards low latitudes, while the expected number of AGNs among unassociated sources is increasing, so that the sum of the two is relatively uniform, as expected for extragalactic sources.

We perform the checks of the classification probabilities using reliability diagrams in Appendix D. We find that the performance of the 3-class classification is similar to the 2-class classification, if we take into account only AGNs and PSRs in the 2-class case. For all associated sources, the 2-class classification overestimates the number of pulsars due to the presence of the other sources, while in the 3-class classification case the other sources are included in the model and the reliability diagrams are reasonably close to the perfect calibration line.

Acknowledgements

The authors would like to thank Jean Ballet, Isabelle Grenier, Pablo Saz Parkinson, and the anonymous referee for valuable comments and suggestions. The work of DM was in part supported by BMBF under the ErUM-Data project "Innovative Digital Technologies for Research on Universe and Matter" (grant number 05H18WERC1) and by DFG grant MA 8279/2-1. We would like to acknowledge the use of the following software: Astropy (`http://www.astropy.org`, Robitaille et al. 2013), matplotlib (Hunter 2007), scikit-learn [`https://scikit-learn.org/stable/about.html`], TOPCAT (Taylor 2005), and Imbalanced-learn Lemaître et al. (2017). This research has also made use of the SIMBAD database, operated at CDS, Strasbourg, France (Wenger et al. 2000).

References

Abdo, A. A., Ackermann, M., Ajello, M., et al. 2010a, ApJS, 188, 405
Abdo, A. A., Ackermann, M., Ajello, M., et al. 2010b, ApJ, 720, 435
Abdollahi, S., Acero, F., Ackermann, M., et al. 2020, ApJS, 247, 33
Acero, F., Ackermann, M., Ajello, M., et al. 2015, ApJS, 218, 23
Ackermann, M., Ajello, M., Albert, A., et al. 2015, J. Cosmology Astropart. Phys., 2015, 008
Ackermann, M., Ajello, M., Albert, A., et al. 2016, Phys. Rev. Lett., 116, 151105
Ackermann, M., Ajello, M., Allafort, A., et al. 2012, ApJ, 753, 83
Ackermann, M., Ajello, M., Baldini, L., et al. 2017, ApJ, 843, 139
Ajello, M., Atwood, W. B., Baldini, L., et al. 2017, ApJS, 232, 18
Ajello, M., Gasparrini, D., Sánchez-Conde, M., et al. 2015, ApJ, 800, L27
Ballet, J., Burnett, T. H., Digel, S. W., & Lott, B. 2020, arXiv e-prints, arXiv:2005.11208
Breiman, L. 2001, Machine Learning, 45, 5
Brewer, B. J., Foreman-Mackey, D., & Hogg, D. W. 2013, AJ, 146, 7
Camilo, F., Kerr, M., Ray, P. S., et al. 2015, The Astrophysical Journal, 810, 85
Caron, S., Dijkstra, K., Eckner, C., et al. 2021, arXiv e-prints, arXiv:2103.11068
Carr, B. J., Kohri, K., Sendouda, Y., & Yokoyama, J. 2010, Phys. Rev. D, 81, 104019
Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. 2002, Journal of Artificial Intelligence Research, 16, 321
Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. 2011, arXiv e-prints, arXiv:1106.1813
Chiaro, G., Salvetti, D., La Mura, G., et al. 2016, MNRAS, 462, 3180
Cox, D. R. 1958, J R Stat Soc B, 20, 215
Daylan, T., Portillo, S. K. N., & Finkbeiner, D. P. 2017, ApJ, 839, 4
Defazio, A., Bach, F., & Lacoste-Julien, S. 2014, arXiv e-prints, arXiv:1407.0202
Di Mauro, M. & Donato, F. 2015, Phys. Rev. D, 91, 123001
Di Mauro, M., Manconi, S., Zechlin, H. S., et al. 2018, ApJ, 856, 106
Doert, M. & Errando, M. 2014, ApJ, 782, 41
Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. 2008, J. Mach. Learn. Res., 9, 1871–1874
Finke, T., Krämer, M., & Manconi, S. 2020, arXiv e-prints, arXiv:2012.05251
Fornasa, M. & Sánchez-Conde, M. A. 2015, Phys. Rep., 598, 1
Friedman, J. H. 2001a, Ann. Statist., 29, 1189
Friedman, J. H. 2001b, Ann. Statist., 29, 1189
H. E. S. S. Collaboration, Abdalla, H., Abramowski, A., et al. 2018, A&A, 612, A1
Hassan, T., Mirabal, N., Contreras, J. L., & Oya, I. 2013, MNRAS, 428, 220
Ho, T. K. 1998, IEEE Transactions on Pattern Analysis and Machine Intelligence, 20, 832
Hogg, D. W. & Lang, D. 2010, in EAS Publications Series, Vol. 45, EAS Publications Series, 351–358
Hopfield, J. 1982, Proc. Nat. Acad. Sci., 79, 2554
Hui, C. Y., Lee, J., Li, K. L., et al. 2020, MNRAS, 495, 1093
Hunter, J. D. 2007, Computing In Science & Engineering, 9, 90
Kingma, D. P. & Ba, J. 2014, arXiv e-prints, arXiv:1412.6980
Kovačević, , M., Chiaro, G., Cutini, S., & Tosti, G. 2019, MNRAS, 490, 4770
Kovačević, M., Chiaro, G., Cutini, S., & Tosti, G. 2020, MNRAS, 493, 1926
Lee, K. J., Guillemot, L., Yue, Y. L., Kramer, M., & Champion, D. J. 2012, MNRAS, 424, 2832
Lefaucheur, J. & Pita, S. 2017, A&A, 602, A86
Lemaître, G., Nogueira, F., & Aridas, C. K. 2017, Journal of Machine Learning Research, 18, 1
Lisanti, M., Mishra-Sharma, S., Necib, L., & Safdi, B. R. 2016, ApJ, 832, 117
Liu, D. C. & Nocedal, J. 1989, Math. Program., 45, 503–528
Liu, W., Bi, X.-J., Lin, S.-J., & Yin, P.-F. 2017, Chinese Physics C, 41, 045104
Luo, S., Leung, A. P., Hui, C. Y., & Li, K. L. 2020, MNRAS, 492, 5377
Malyshev, D. & Hogg, D. W. 2011, ApJ, 738, 181
Mirabal, N., Charles, E., Ferrara, E. C., et al. 2016, ApJ, 825, 69

Miville-Deschênes, M.-A., Murray, N., & Lee, E. J. 2017, ApJ, 834, 57
Nolan, P. L., Abdo, A. A., Ackermann, M., et al. 2012, ApJS, 199, 31
Peretto, N., Lenfestey, C., Fuller, G. A., et al. 2016, A&A, 590, A72
Peron, G., Aharonian, F., Casanova, S., Zanin, R., & Romoli, C. 2020, ApJ, 896, L23
Ritchey, A. M. 2020, MNRAS, 495, 2909
Robitaille, T. P., Meade, M. R., Babler, B. L., et al. 2008, AJ, 136, 2413
Robitaille, T. P., Tollerud, E. J., Greenfield, P., et al. 2013, A&A, 558, A33
Saha, L., Domínguez, A., Tibaldo, L., et al. 2020, ApJ, 897, 131
Salvetti, D., Chiaro, G., La Mura, G., & Thompson, D. J. 2017, MNRAS, 470, 1291
Saral, G., Hora, J. L., Audard, M., et al. 2017, ApJ, 839, 108
Saz Parkinson, P. M., Xu, H., Yu, P. L. H., et al. 2016, ApJ, 820, 8
Schmidt, M., Le Roux, N., & Bach, F. 2017, Math. Program., 162, 83–112
Simpson, R. J., Povich, M. S., Kendrew, S., et al. 2012, MNRAS, 424, 2442
SOM. 2021, The probabilistic catalogs are available online in the SOM folder of the paper source: `https://arxiv.org/format/2102.07642`
Taylor, M. B. 2005, in Astronomical Society of the Pacific Conference Series, Vol. 347, Astronomical Data Analysis Software and Systems XIV, ed. P. Shopbell, M. Britton, & R. Ebert, 29
Wenger, M., Ochsenbein, F., Egret, D., et al. 2000, A&AS, 143, 9
Yang, R.-Z. & Wang, Y. 2020, A&A, 640, A60
Zechlin, H.-S., Cuoco, A., Donato, F., Fornengo, N., & Regis, M. 2016a, ApJ, 826, L31
Zechlin, H.-S., Cuoco, A., Donato, F., Fornengo, N., & Vittino, A. 2016b, ApJS, 225, 18
Zhu, K.-R., Kang, S.-J., & Zheng, Y.-G. 2021, Research in Astronomy and Astrophysics, 21, 015

Appendix A: Tests of additional meta-parameters

In this appendix we discuss tests of some meta-parameters, which had a relatively little effect on the accuracy of the algorithms. For these tests we use the 2-class classification in the 3FGL catalog.

For the LR algorithm, we test two additional meta-parameters: regularization and tolerance. The effect of the choice of these parameters on accuracy is less than 1% (Fig. A.1). Therefore we used the default values for these parameters (tolerance is 10^{-4} and regularization parameter is 1) both in the 2-class and in the 3-class classifications.

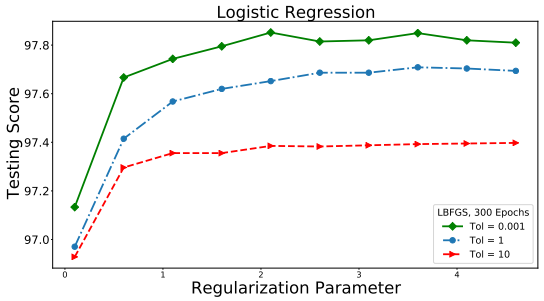

Fig. A.1. Dependence of LR on tolerance and regularization.

In Fig. A.2 we show the effect of adding the second hidden layer in the NN algorithm. The difference between the best accuracies of the NN with one hidden layer (cf. Table 2) and the NN with the additional hidden layer is less then 1%.

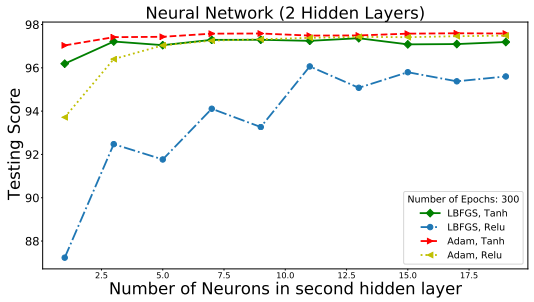

Fig. A.2. Dependence of NN on the number of neurons in the second hidden layer, for 11 neurons in the first hidden layer.

We summarize features and their statistics, which we use for probabilistic classification of sources in the 3FGL and 4FGL-DR2 catalogs in Tables A.1 and A.2 respectively. We show the feature importances for the 2-class classification of 4FGL-DR2 sources in Table A.3. Similarly to feature importances for the 3FGL 2-class classification reported in Table 1, some of the most important features are significance of curvature in the spectrum, variability index, energy flux above 100 MeV and its uncertainty.

Feature Name	Mean	Standard Deviation	Minimum	Maximum
GLON	182.58	100.9	0.59	359.64
GLAT	2.67	41.1	−87.66	86.37
ln(Energy_Flux100)	−25.34	1.05	−27.13	−18.53
ln(Unc_Energy_Flux100)	−27.47	0.47	−28.47	−24.78
ln(Signif_Curve)	0.23	1.19	−5.81	4.44
ln(Variability_Index)	4.35	0.95	3	11.01
500MeV_Index	2.16	0.37	−1.61	3.75
HR12	−0.41	0.5	−1	1
HR23	−0.53	0.36	−1	1
HR34	−0.55	0.25	−1	1
HR45	−0.59	0.33	−1	1

Table A.1. Statistics of features used for 2 class probabilistic classification of the 3FGL sources.

Feature Name	Mean	Standard Deviation	Minimum	Maximum
GLON	179.65	101.91	0.09	359.99
GLAT	1.93	41.02	−87.68	87.57
ln(Energy_Flux100)	−26.08	1.11	−28.5	−18.48
ln(Unc_Energy_Flux100)	−28.19	0.57	−29.67	−24.25
ln(Pivot_Energy)	7.45	0.78	5.01	10.09
LP_Index	2.14	0.38	−0.08	5.31
Unc_LP_Index	0.16	0.13	0	3.55
LP_beta	0.16	0.2	−0.17	1
LP_SigCurv	2.89	6.99	0	225.77
ln(Variability_Index)	3.14	1.4	−0.81	11.23
HR12	0.12	0.75	−1	1
HR23	−0.26	0.61	−1	1
HR34	−0.52	0.36	−1	1
HR45	−0.54	0.26	−1	1
HR56	−0.66	0.28	−1	1
HR67	−0.56	0.52	−1	1

Table A.2. Statistics of features used for 2 class probabilistic classification of the 4FGL-DR2 sources.

Feature	RF: 50, 6	BDT: 100, 2
ln(LP_SigCurv)	0.297	0.465
LP_beta	0.151	0.109
ln(Variability_Index)	0.085	0.253
ln(Unc_Energy_Flux100)	0.081	0.059
ln(Energy_Flux100)	0.076	0.008
HR56	0.071	0.015
Unc_LP_Index	0.067	0.009
HR34	0.035	0.005
ln(Pivot_Energy)	0.031	0.006
HR23	0.025	0.005
LP_Index	0.015	0.016
HR67	0.015	0.010
HR45	0.015	0.006
GLON	0.013	0.017
HR12	0.009	0.004
GLAT	0.007	0.003

Table A.3. Feature importances for classification of 4FGL-DR2 sources using RF (50 trees, max depth 6) and BDT (100 trees, max depth 2) algorithms ordered by decreasing importance in the case of RF algorithm.

Appendix B: Comparison of Oversampling methods

In this appendix we compare the method of oversampling by repeating sources described in Section 3.3 with SMOTE (Chawla et al. 2011). We use SMOTE from the Imbalanced-learn library, which is based on an implementation of Chawla et al. (2002). We estimate the probabilities of classification of all sources in the 3FGL and 4FGL-DR2 catalogs using the same algorithms and meta-parameters as in Section 3.3. The only difference

Classes	2-Class		3-Class	
Method	Mean	Std.	Mean	Std.
RF	-0.431	0.461	0.114	0.343
BDT	-0.161	0.426	-0.102	0.337
NN	0.285	0.419	0.224	0.254
LR	0.437	0.615	0.357	0.325

Table B.1. The parameters of the distribution of Δ in Eq. (B.1) for the 4 algorithms used in 3FGL with PSR-like probabilities for all sources.

Classes	2-Class		3-Class	
Method	Mean	Std.	Mean	Std.
RF	-0.335	0.489	-0.491	0.464
BDT	-0.120	0.581	0.007	0.423
NN	0.472	0.430	0.171	0.145
LR	0.531	0.732	0.347	370

Table B.2. The parameters of the distribution of Δ in Eq. (B.1) for the 4 algorithms used in 4FGL-DR2 with PSR-like probabilities for all sources.

in the SMOTE case is the oversampling technique. First, we compare the difference in class probabilities for individual sources relative to the statistical uncertainty due to the random choice of the training samples. In particular, we calculate the difference of probabilities to belong to, e.g., PSR class relative to the maximal standard deviation between the oversampling-by-repeating and SMOTE

$$\Delta = \frac{P_O - P_S}{max(\sigma_O, \sigma_S)}, \tag{B.1}$$

where P_O (P_S) is the probability for the oversampling-by-repeating (SMOTE) and σ_O (σ_S) is the corresponding standard deviation. The mean and the standard deviation of Δ for 2- and 3-class cases in the 3FGL and 4FGL-DR2 catalogs are presented in Tables B.1 and B.2 respectively. We note that in the 3-class case we use less oversampling than in the 2-class case, namely, the oversampling factor in the 3-class case is equal to the square root of the ratio of the number of sources, while in the 2-class case it is equal to the ratio of the number of associated AGNs to the number of associated pulsars. This is the reason for the smaller bias and standard deviation of the difference in the 3-class case relative to the 2-class case. Overall, the differences for individual probabilities are smaller than the uncertainties due to randomness of training.

The LR algorithm has some of the largest differences. We plot the histogram of the Δ for the PSR-like probabilities in the 3FGL (4FGL-DR2) catalogs in Fig. B.1 (B.2).

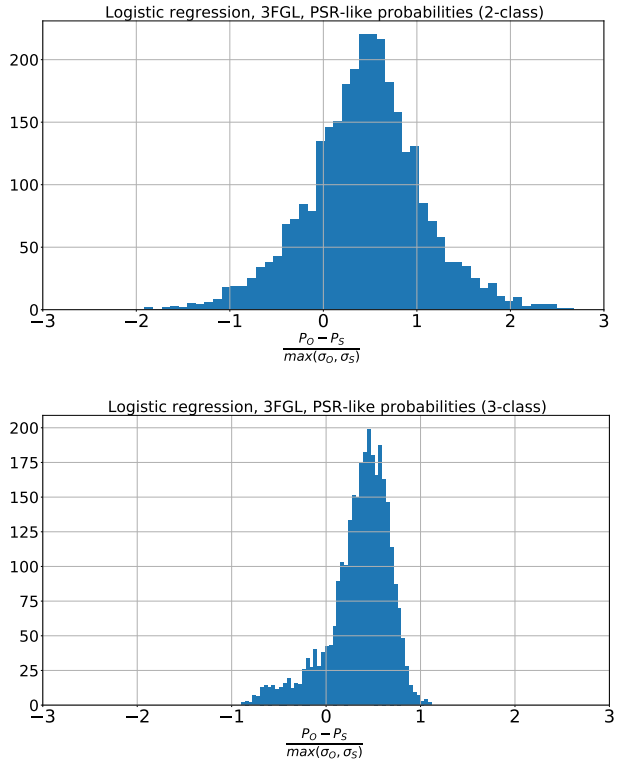

Fig. B.1. Distribution of difference of probabilities derived with oversampling-by-repeating (P_O) and SMOTE (P_S) relative to the standard deviations due to random choice of training samples for LR in the 2-class (top) and the 3-class (bottom) classifications of the 3FGL sources.

The difference between oversampling-by-repeating and SMOTE is illustrated in Fig. B.3. The domains are determined by averaging over 100 random choices of the training data. One of these choices is shown on the plots: in this case, both oversampling-by-repeating and SMOTE have the same training and testing samples. In the oversampling-by-repeating the training sources are oversampled by simply repeating the sources, while in SMOTE new sources are created by randomly placing sources in the parameter space along the lines connecting a source and one of its nearest neighbours. In our implementation we choose one out of the 5 nearest neighbours.

Although the mean and the standard deviations for the probabilities of the individual sources are smaller than the statistical uncertainties of the probabilities, the presence of the bias can have a significant effect when we sum the probabilities, e.g., in population studies. In order to check this effect we compare the source count distributions as a function of flux for oversampling-by-repeating and SMOTE in Fig. B.4. We show the expected number of pulsars among unassociated sources for the 2- and 3-class cases using the 3FGL and 4FGL-DR2 catalogs. We also show the expected number of OTHER sources in the 3-class case. The difference can indeed be significant for some of the algorithms, but the change is compara-

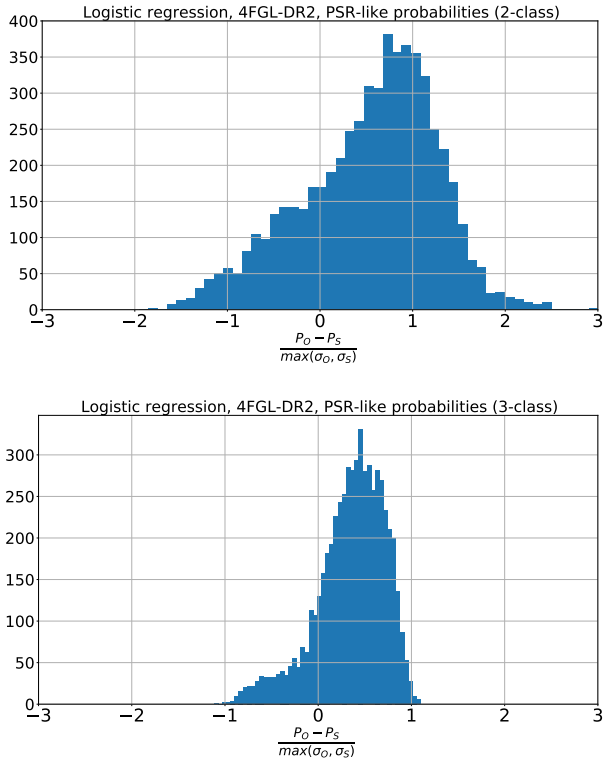

Fig. B.2. Same as Fig. B.1 for the 4FGL-DR2 catalog.

ble or smaller than the difference among the algorithms. The classification probabilities in the 2- and 3-class cases with SMOTE for the 3FGL and 4FGL-DR2 catalogs are available online in the SOM/SMOTE subfolder in the supplementary online materials (SOM 2021).

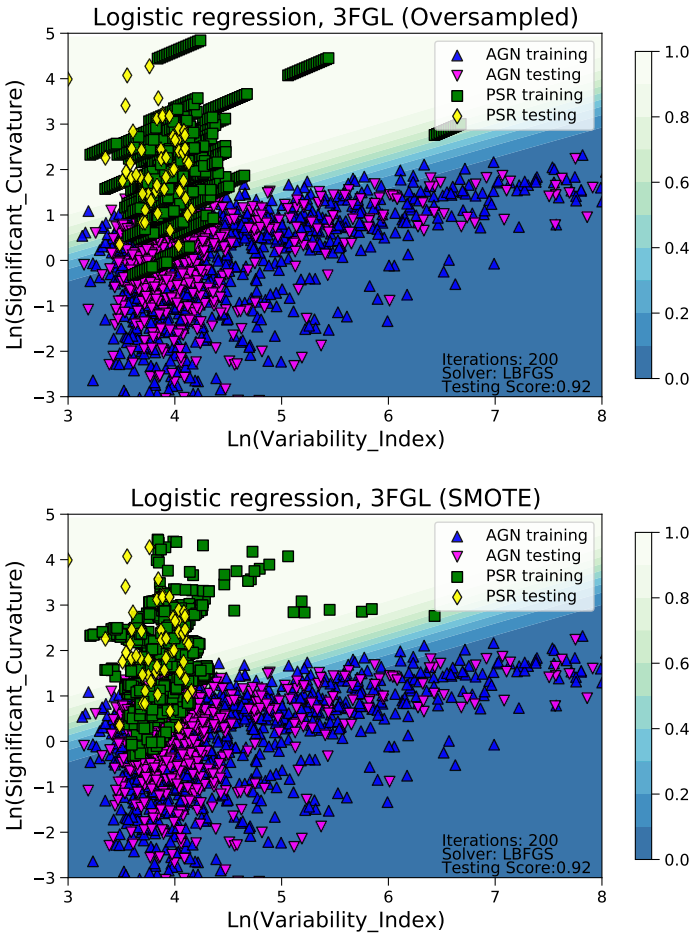

Fig. B.3. Comparison of classification domains for the 2-class classification of the 3FGL sources using LR with oversampling by repeating sources (top) and SMOTE (bottom).

Appendix C: Choice of the classification threshold

In the paper we use the largest probability rule to classify sources by individual algorithms. In particular, this means that in the 2-class case a source is classified as an AGN or a pulsar if the corresponding probability is larger than 50%. Similarly, in the 3-class case a source may be classified based on a minimum required 33.3% probability. Classification by the largest probability is a common approach in machine learning and we use it for the calculation of accuracy of individual algorithms and for the optimization of the meta-parameters. However, one can choose other thresholds for the classification depending on the goals of the analysis. For instance, if a list of high-probability pulsars is required one could use a higher threshold, which is expected to reject most of the false candidates but real pulsars would be rejected as well. In general, one expects a higher precision for higher threshold at the expense of lower recall. Lower threshold would increase the recall (one would miss fewer true candidates), but decrease the precision (there will be more false positives).

In this appendix we study the effect of changing the threshold for the probabilistic classification of sources by the individual algorithms on the overall precision and recall of classification using the agreement of all 8 algorithms. In Figure C.1 we show precision and recall for the 2- and 3-class classification of PSR and OTHER sources in the 3FGL and 4FGL-DR2 catalogs. In the calculation we use the probabilistic classification of associated sources described in Section 4.1, i.e., we perform 1000 random splits into training and testing samples and determine the class probabilities for a source by averaging the probabilities when the source is included in the test samples. We note that with increasing threshold the recall is decreasing while the precision is generally increasing (except for a few high threshold values in the 3FGL PSR and OTHER 3-class classification, where the number of candidates is very small). The precision in the 3-class case is generally better than in the 2-class case (see the top panels of Figure C.1), while the recall is better in the 2-class case.

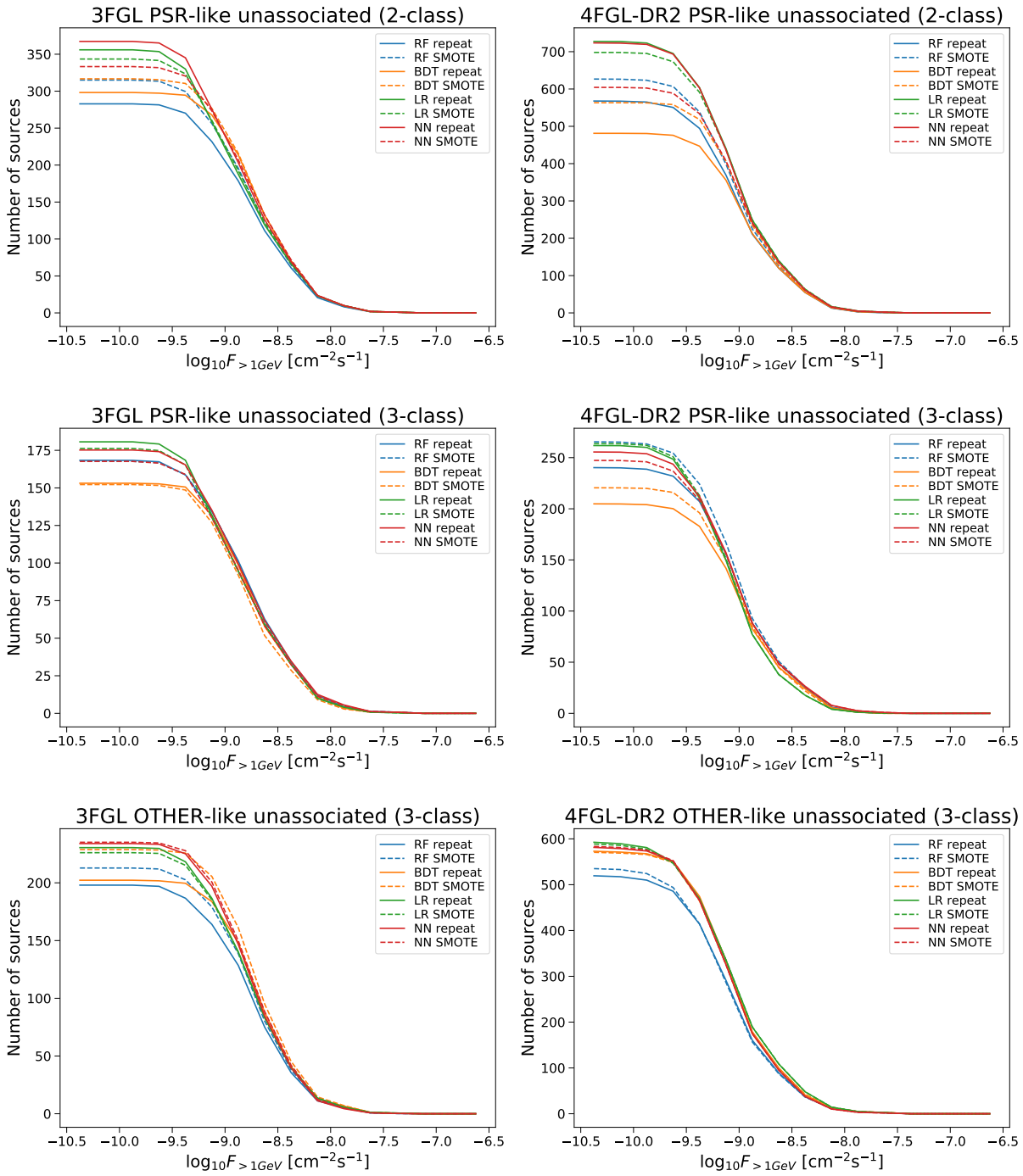

Fig. B.4. Comparison of source count distributions as a function of flux for the estimated number of PSR and OTHER sources among the unassociated sources in the 2- and 3-class cases for oversampling-by-repeating and SMOTE. The number of PSR-like sources in the 2-class case is not corrected for the presence of OTHER sources. The changes due to using SMOTE oversampling technique are comparable to the difference among the algorithms for oversampling by repeating.

There are a few points for high threshold values where no associated 3FGL sources were classified as OTHER. In this case the precision is undetermined due to division by zero and the corresponding points are absent in the precision curve on the lower left panel of Figure C.1.

In Figure C.2 we show the precision and recall for different thresholds for classification of unassociated sources in the 3FGL catalog, which have associations in the 4FGL-DR2 catalog. In general the precision and recall in this case are smaller than the estimates in Figure C.1. The estimates in Figure C.2 are likely more realistic than in the test samples case, since they also take into account possible difficulties in reconstructing the properties of the sources, such as the spectrum, which can affect the probabilistic classification of the sources.

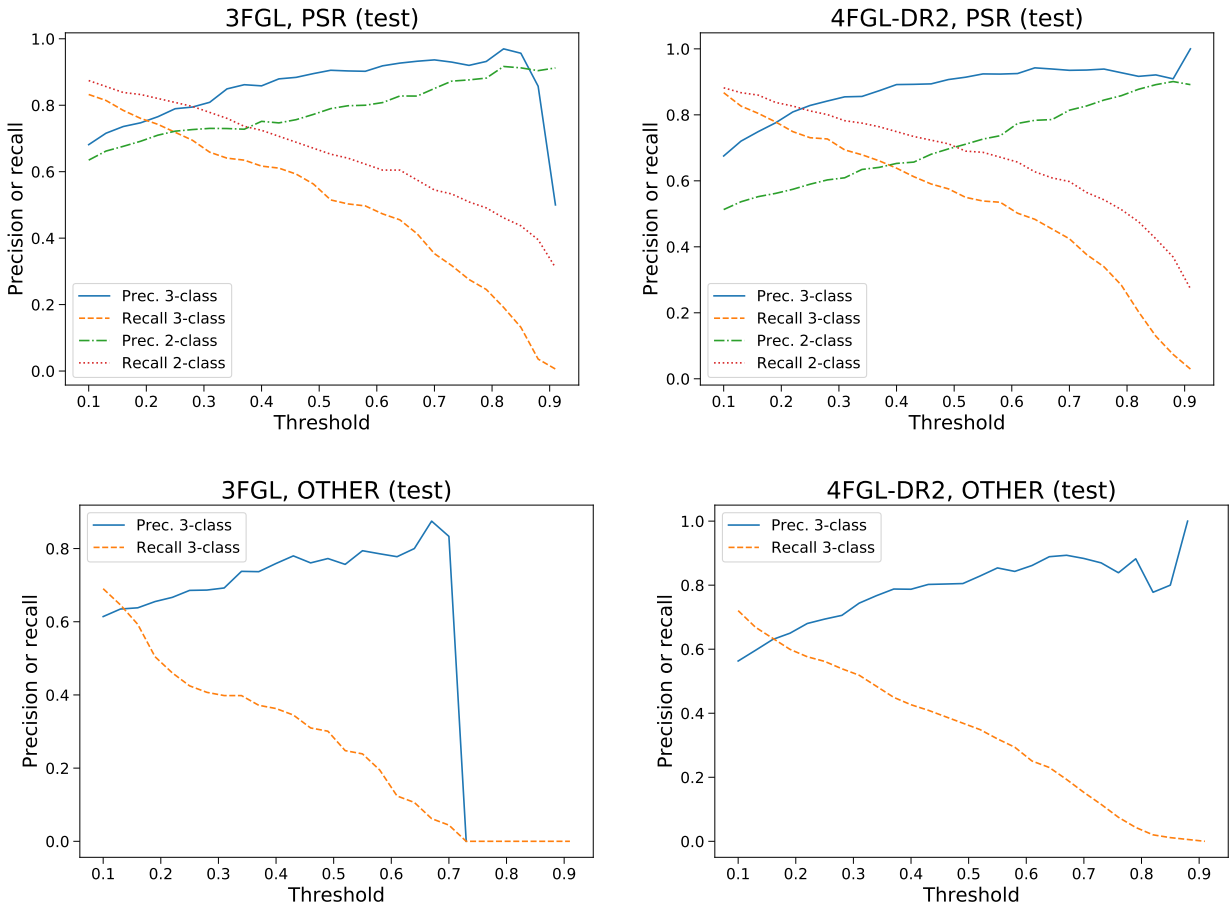

Fig. C.1. Precision and recall for different choices of the threshold for classification of sources by individual algorithms. The all-algorithms-agree method is used for the final classification. In these estimates, we use associated sources, the corresponding class probabilities are calculated by averaging over the probabilities when the sources are included in testing samples. Above the threshold of 0.73, no sources were classified as OTHER in the 3-class classification of the 3FGL sources. The corresponding points in the precision curve on the bottom left panel are absent.

Appendix D: Reliability diagrams

One of the characteristics of a classification algorithm is the reliability diagram (also known as the calibration curve), where one compares predicted probabilities with the true fractions of correct classifications. For the calculations we use "calibration_curve" function implemented in scikit-learn.

In Fig. D.1 we show the reliability diagram for PSR classification in the 2- and 3-class cases for the 3FGL and 4FGL-DR2 catalogs. The predicted PSR-like probabilities for the associated sources are separated in 10 equally spaced bins between 0 and 1, the x-values are the average predicted probabilities in each bin, the y-values are the fractions of associated pulsars among the sources with predicted probabilities in the corresponding bins. The solid lines at 45° show the perfect calibration, when the expected probabilities are on average equal to the fraction of true classifications. If the curve is above (below) the perfect calibration line, then the expected probabilities are smaller (larger) than the true fraction, i.e., the algorithm underestimates (overestimates) the true fraction.

In the top panels of Fig. D.1 we show the reliability diagrams for the 2-class classification where we take into account only sources in PSR and AGN classes. One can see that without oversampling, some algorithms tend to underestimate the true fraction (e.g., RF and LR in the 3FGL 2-class case), while with oversampling, the algorithms generally overestimate the true fraction of PSRs (e.g., NN_O and LR_O for 3FGL and all oversampling algorithms for 4FGL-DR2) – this behavior is not unexpected, since in the oversampling case we artificially increase the number of sources in the smaller class.

In the middle panels of Fig. D.1 we show the reliability diagrams for the 2-class classification when we add the OTHER sources. In this case all algorithms underestimate the true fraction of PSRs, this is due to the presence of additional sources, none of which are pulsars. It shows that the 2-class classification is likely overestimating the number of pulsars among the unassociated

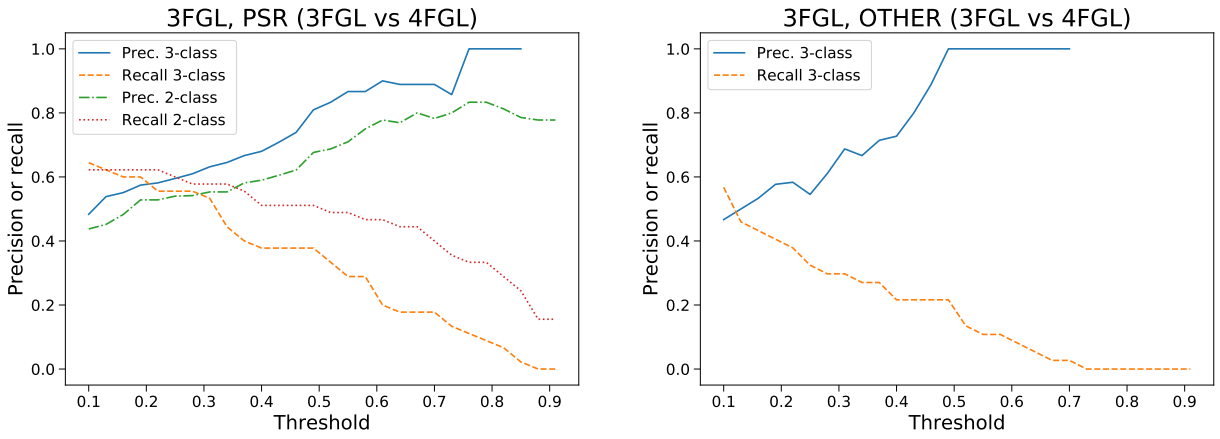

Fig. C.2. Precision and recall for different choices of the threshold for classification of sources by individual algorithms. The all-algorithms-agree method is used for the final classification. The precision and recall are calculated for unassociated 3FGL sources, which have associations in the 4FGL-DR2 catalog. The classes in the 4FGL-DR2 catalog are considered as the true classes. The points on some precision curves are absent for high thresholds, when there are no sources classified as pulsars (OTHER) on the left (right) panel in the 3-class case.

sources due to the presence of the OTHER sources. Thus some correction or calibration is needed.

The bottom panels of Fig. D.1 show the reliability diagrams in the 3-class case. One can see that the performance of the algorithms is not worse than in the 2-class case on the top panels (the better performance of the oversampling cases can be in part attributed to fewer sources added in the oversampling for the 3-class case). In Fig. D.2 we show the reliability diagrams for the OTHER class in the 3-class classification. One can see that the performance of the algorithms is also good in this case, although the OTHER class is the smallest class and has different types of sources, which can in principle lead to confusion with other classes and poor performance of classification.

Overall, we find that the reliability of predictions in the 3-class case is similar to the performance in the 2-class case when only PSR and AGN classes are taken into account. In addition, we expect a similar performance for the unassociated sources, since the OTHER class is included in the calculation of the reliability diagrams, i.e., no correction is needed as in the 2-class case. We also note, that although a particular algorithm can be above or below the perfect calibration curve, the envelope of the predictions contains the perfect calibration curve, i.e., the envelope of the predictions gives a reasonable estimate of the modeling uncertainty.

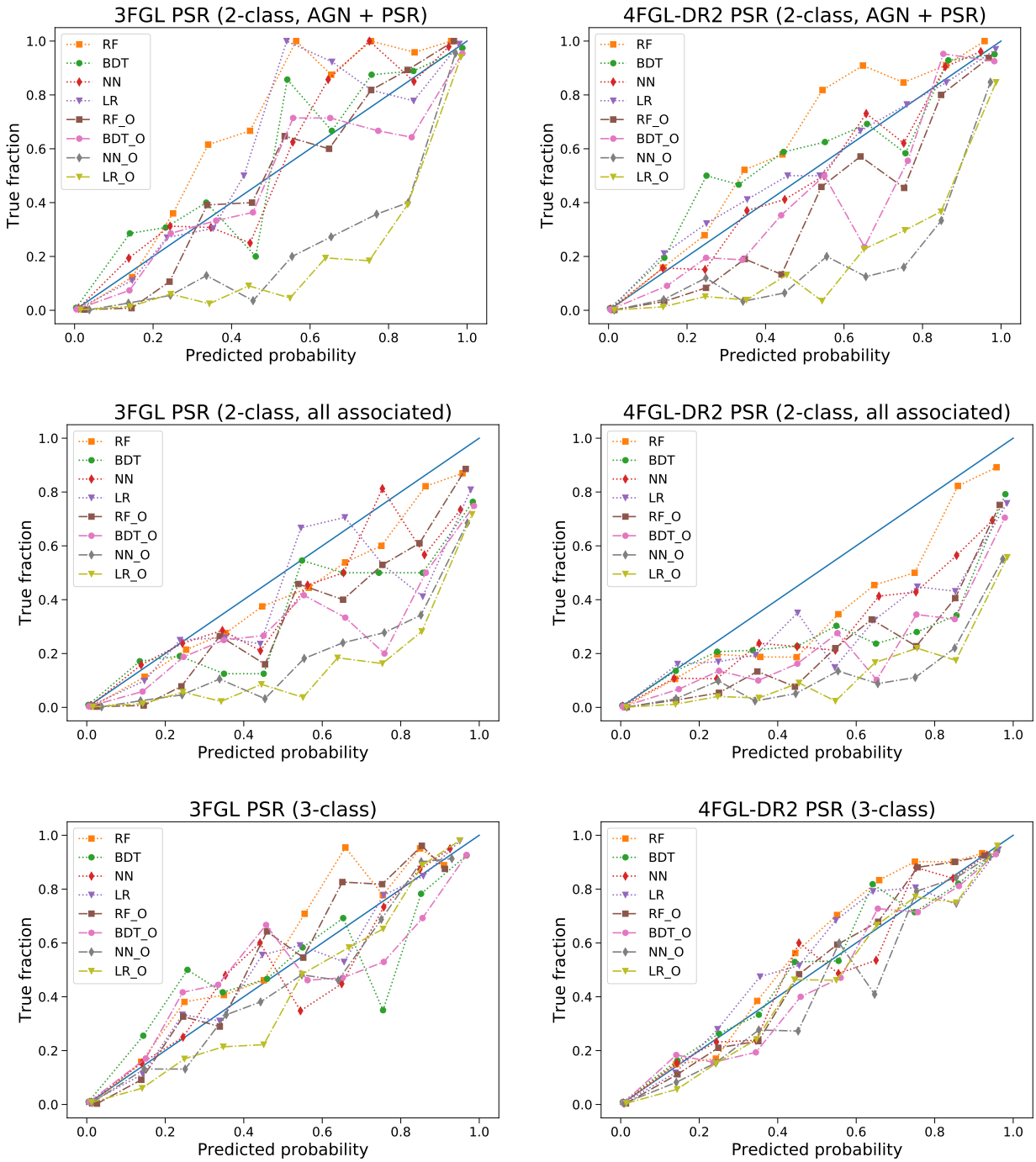

Fig. D.1. Reliability diagrams for the PSR class. Top panels: 2-class classification taking into account only AGN and PSR associated 3FGL and 4FGL-DR2 sources. Middle panels: same as the top panels but taking into account all associated 3FGL and 4FGL-DR2 sources. Bottom panels: 3-class classification of 3FGL and 4FGL-DR2 sources.

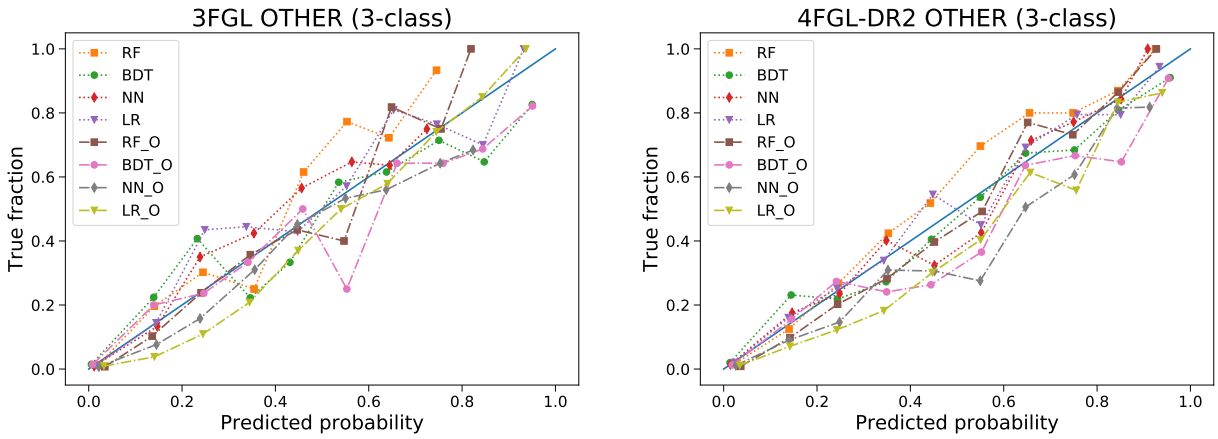

Fig. D.2. Reliability diagrams for the OTHER class in the case of the 3-class classification of the 3FGL and 4FGL-DR2 sources.