Astronomy & Astrophysics manuscript no. 0head_ML_3FGL ©ESO 2020
December 11, 2020

Machine learning methods for probabilistic *Fermi*-LAT catalogs

A. Bhat [⋆1] and D. Malyshev [⋆⋆1]

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

Received September 15, 1996; accepted March 16, 1997

ABSTRACT

Context. Classification of sources is one of the most important tasks in astronomy. Sources detected in one wavelength band, e.g., in 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 *Fermi* Large Area Telescope (LAT) point source catalogs (3FGL and 4FGL) into two major classes: pulsars and active galactic nuclei (AGNs).

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

Results. We determine probabilistic classification of both associated and unassociated sources in 3FGL and 4FGL catalogs. We find the accuracy of classification to be about 94% – 97% (96% – 98%) by splitting associated 3FGL (4FGL) sources into training and testing samples. We cross-check the accuracy by comparing the predicted classes of unassociated sources in 3FGL that have associations in 4FGL, the corresponding accuracy is 87% – 96% depending on the ML algorithm. We find that most of unassociated sources are likely AGNs, while the expected number of pulsars among unassociated sources in the 3FGL (4FGL) catalog is 267 ± 110 (386 ± 179), i.e., the total number of pulsars is expected to be about two times larger than the number of associated pulsars: 167 (239) in the 3FGL (4FGL) catalogs.

Key words. Methods: statistical – Catalogs

Contents

[⋆] 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.mayshev@fau.de

1. Introduction

Multiwavelength association of astronomical sources is important for understanding their nature. Unfortunately, in many cases a firm association of sources at different wavelength is not possible. For example, about one third of 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 useful to know at least to which class the unassociated sources belong to or, which is more typical, what are the probabilities for the source to belong to various classes. In this paper we use several machine learning (ML) algorithms to find probabilistic classification of unassociated sources in the *Fermi*-LAT third (3FGL) and fourth (4FGL) catalogs (Acero et al. 2015; Abdollahi et al. 2020). We will refer to the catalogs, where the classification of the sources is given by 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 or that a source is an overlay of two sources etc. All these possibilities can be included in the probabilistic catalogs, which were previously 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 introduced in the 4FGL catalog (Abdollahi et al. 2020) for faint sources. Probabilistic classification of unassociated *Fermi*-LAT sources was performed by, e.g., Ackermann et al. (2012); Saz Parkinson et al. (2016); Mirabal et al. (2016); Lefaucheur & Pita (2017); Luo et al. (2020); Zhu et al. (2020), or in the application for subclassification 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 classification of gamma-ray sources into two classes: AGNs and pulsars. We revisit probabilistic classification of 3FGL sources and compare results of classification of unassociated sources with associations in 4FGL. We also determine probabilistic classification of the 4FGL sources. We calculate the expected number of pulsars and AGNs among unassociated sources in 3FGL and 4FGL catalogs including correction for the presence of other sources, neither pulsars nor AGNs, among the unassociated sources.

Catalogs of gamma-ray point sources are typically designed to have low false detection rate. Nevertheless, 469 sources out of 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 3FGL without counterparts in 4FGL 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 an 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, lower threshold would lead to more false detections due to fluctuations of the background. Thus, on the one hand, 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, 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 are probabilistically classified into classes, which include the statistical fluctuations class. At low confidence, the probability for a source to come from a background fluctuation is high. This probability is decreasing as the significance of sources increases. Apart from the statistical fluctuations class, classes can include various types of Galactic and extra-galactic sources, diffuse emission deficiencies, extended sources. A user of such a catalog will have 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 probabilistic classification of *Fermi*-LAT sources into AGNs and pulsars. 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 of constructing the probabilistic catalogs and the choice of the ML methods. In Section 3 we construct the classification algorithms using the associated source in the 3FGL catalog for training. We consider several aspects: 1) feature selection, 2) training of the algorithms and selection of the 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 sources. We compare the predictions for the unassociated sources in 3FGL with the associations in 4FGL. We also retrain the algorithms with the same meta-parameters as in the 3FGL case using associated 4FGL sources and construct a probabilistic catalog based on 4FGL. In Section 5 we show applications of the probabilistic catalogs for predicting the number of pulsars and AGNs among the unassociated source 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 catalogs. In Section 6 we present conclusions.

2. Choice of methods

2.1. General methodology

The first choice one has to make in constructing a probabilistic catalog is the input data and the machine learning methods. For the input data we take associated PS in the 3FGL catalog, which we 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, instead of selecting the best one. The difference among the predictions will serve as a measure of modeling uncertainty related to the choice of the classification algorithm.

2.2. Discussion of the choice 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 in mostly Galactic and extragalactic by a cut on the Galactic latitude. At the next step, the high latitude sources can be further subsplit 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 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 too shallow trees would not be able to describe the data well, which can lead to bias. As a result, one needs to be very careful in 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 is that it can miss the classification of some subsets of data. In BDT algorithms, 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 of 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 features. In particular, for two-class classification the probability of class 1, given the set of features x, is modeled by sigmoid (logit) function

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

The probability of class 0 is then modeled as $p_0(x) = 1 - p_1(x)$. 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 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 Equation (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 corresponds 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. Neural network with several hidden layers is obtained by a sequence of nonlinear 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 etc. In the context of neural networks, the non-linear transformations are called activation functions. If the activation function for the output layer is sigmoid, then the output value (values) can be interpreted as probabilities.

3. Construction of a probabilistic catalog

As an example of the construction of a probabilistic catalog, we will use the 3FGL catalog. For training and testing the methods, we use sources which have associations and no missing values in the catalog. In this paper we will 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 either a pulsar or an AGN. After the training of the algorithms, we test the performance with the test sources and predict the classes of unassociated sources that have all features present in the catalog table. The general workflow will have the following steps:

1. Select data for learning 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 prediction for unassociated point sources of the 3FGL. We also apply the classification for 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 use in the following section for a construction of probabilistic catalogs based on the *Fermi*-LAT 3FGL and 4FGL catalogs.

3.1. Data and feature selection

We restrict the analysis to associated and unassociated sources without any missing or statistically insignificant values (e.g., none or infinity). We use the associated sources 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), which results in a list of 1905 sources.

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. In the first part of our work, we use the following ten features: ln(Flux_Density), ln(Unc_Energy_Flux100), Spectral_Index, ln(Signif_Curve), four hardness ratios $hr_{ij} = \frac{EF_j - EF_i}{EF_j + EF_i}$, where EF_i is the energy flux in bin i, ln(Variability_Index), and the galactic latitude GLAT. The table of features and their statistics can be found in the appendix.

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 the 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). We will split the data randomly into 70% training and 30% testing samples.

3.2.1. Random Forests

The two main parameters characterizing the random forest 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).

Figure 1 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

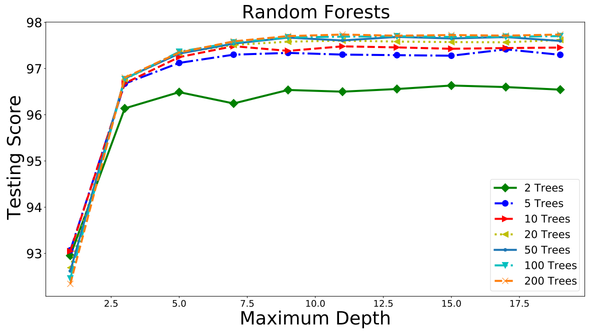

Fig. 1. Test score (accuracy) of RF classification as a function of the number of trees and the maximal depth of trees.

number of allowed features is $\sqrt{\# \text{ features}}$). It is also insensitive to the number of trees above approximately 20 trees. For classification we will use 50 trees with the maximum depth of 6.

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 feature importances for the case of two different RF algorithms: 20 trees with maximum depth of 2 and 50 trees with maximum depth of 6 are shown in Table 1. We find that the three most important features for both cases are significance of curvature, hardness ratio of the last two energy bins, and spectral index (shown in bold font in the table).

Feature	20 trees, depth 2	50 trees, depth 6
ln(Flux_Density)	0.04	0.05
ln(Unc_Energy_Flux100)	0.07	0.08
Spectral_Index	0.13	0.11
ln(Signif_Curve)	0.33	0.29
ln(Variability_Index)	0.06	0.10
hr12	0.03	0.03
hr23	0.03	0.04
hr34	0.02	0.03
hr45	0.25	0.21
GLAT	0.01	0.02

Table 1. Feature importances for two RF algorithms.

It is interesting to note that Galactic latitude is among the least significant features. We have also used sine of 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 5.2, where we calculate the latitude and longitude profiles of the associated and unassociated source counts.

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 spectral index, and plot the resulting probabilities of classes in Figure 2. We compare classification domains for two models: 20 trees with the maximum depth of 2 and 50 trees with the maximum depth of 6. The depth 6 model has a better testing accuracy than the depth 2 model due to the more nuanced probability distribution. It also illustrates how more complex algorithms can over-fit the data (notice the narrow bands in probability distribution for the maximum depth of 6 case), although overfitting is not an issue for RF algorithms in our case. It is important to note that this plot is for one run only (we split the data into training and testing samples only one time) and the model is trained on two features, whereas in the final classification with RF, we use 10 features and average over 1000 splits into training and testing samples in a model with 50 trees with a maximum depth of 6.

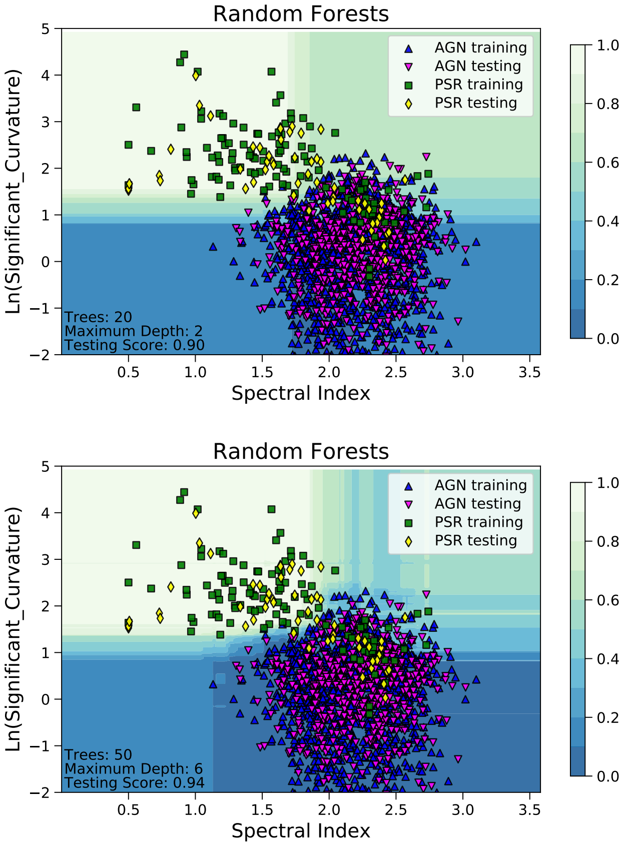

Fig. 2. RF classification domains showing class probabilities for training with two features. We compare RF with 20 trees and maximum depth of 2 and 50 trees and maximum depth of 6.

3.2.2. Boosted Decision Trees

The meta-parameters for BDT algorithms are similar to RF algorithms: they are the number of trees and the maximal 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 Figure 3. Unlike the RF, which is also an ensemble based method, the testing accuracy drops for the maximal depths larger than 2.

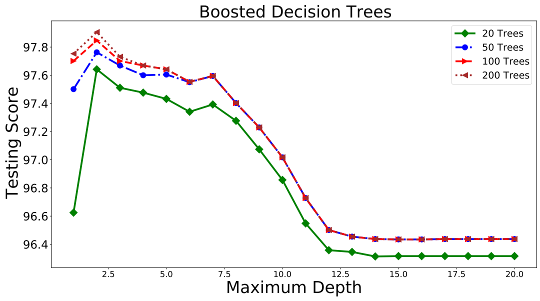

Fig. 3. Dependence of BDT accuracy on maximum depth and the numbers of trees.

The classification domains in case of two features for the maximum depth of 2 and 12 are presented in Figure 4. It is interesting to note that already for the maximum depth of 2, the domains show a sign of overfitting: narrow bands in the class probabilities. For the classification we will use BDT with 100 trees and the maximum depth of 2.

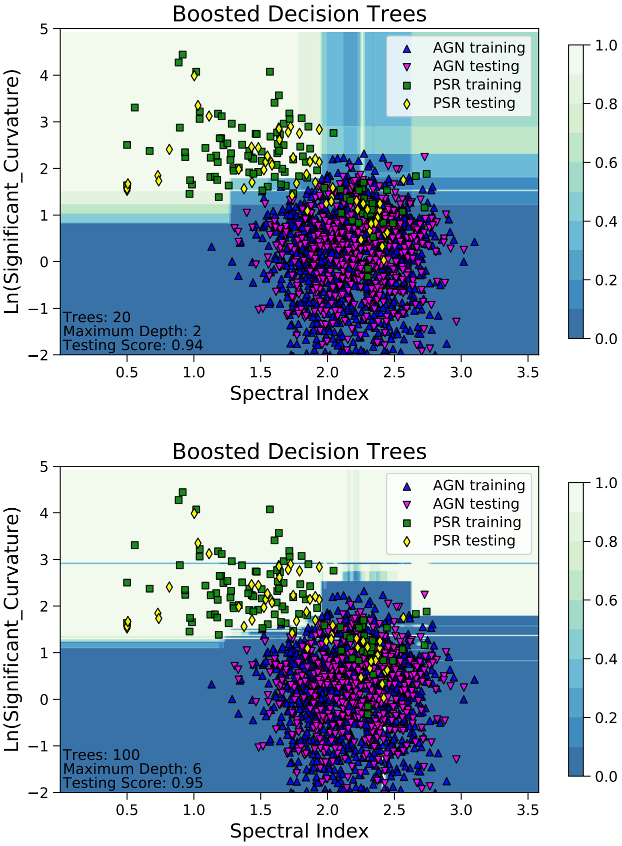

Fig. 4. Classification domains for BDT for training with two features.

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.

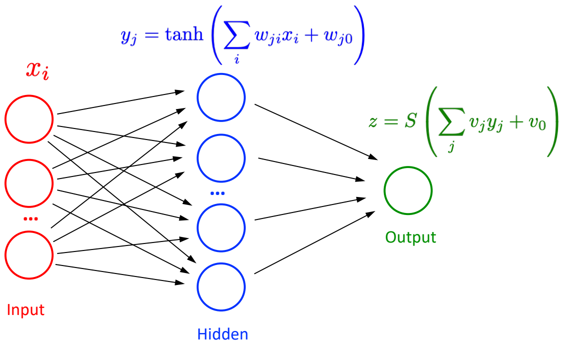
$y_j = \tanh\left(\sum_i w_{ji}x_i + w_{j0}\right)$
x_i
$z = S\left(\sum_j v_j y_j + v_0\right)$
...
...
Input
Hidden
Output

Fig. 5. 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 Figure 5. It is a fully connected NN with 10 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. This is the same loss function as in the LR case. 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 accuracy on the number of neurons in the hidden layer, on the activation function, and on the optimization algorithm is shown in Figure 6. 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 train-

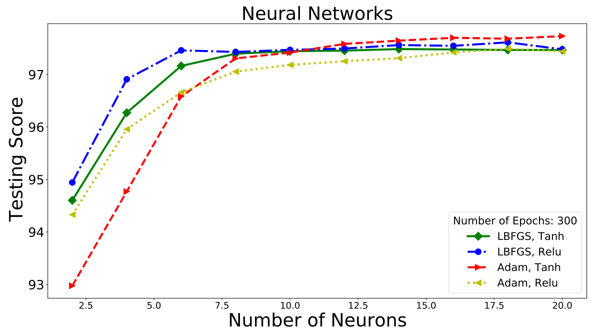

Fig. 6. Dependence of accuracy on the number of neurons for different NN models.

ing. About 10 neurons in 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 Figure 7. The accuracy increases with higher number of epochs and saturates at around 50 for LBFGS and 300 for Adam. LBFGS converges faster than Adam, which uses stochastic gradient descent.

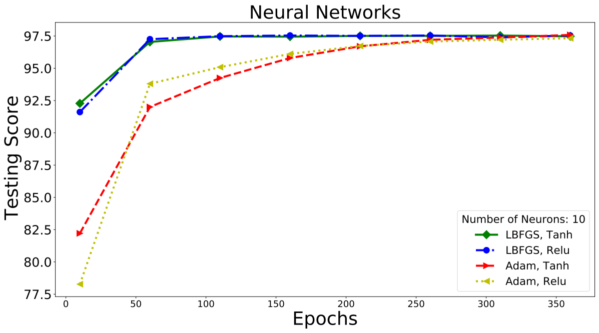

Fig. 7. Dependence of testing accuracy on the number of epochs in training for different solvers and activation functions.

We illustrate the classification domains for NN with two input features in Figure 8. In this case we also use only two neurons in the hidden layer. The figure on the top panel shows the domains after 50 epochs, which has a much less determined boundary than the domains in the bottom panel derived for 300 epochs. One can also see that the separation boundary is smoother compared to the RF domains in Figure 2 or BDT domains in Figure 4.

For our final model we chose one hidden layer with ten neurons, 300 training epochs, Adam solver, and tanh activation function at the hidden layer.

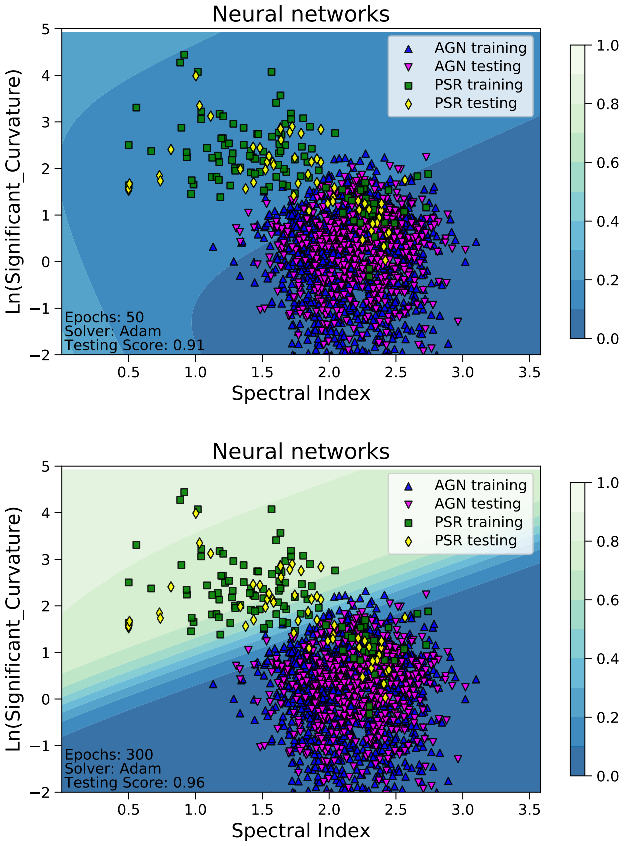

Fig. 8. NN classification domains for 2 input features. We use 2 neurons in the hidden layer. We use tanh activation function and Adam solver. Top panel: 50 training epochs, bottom panel: 300 training epochs.

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 $m(x) = \sum_{k=1}^{10} f_k x_k$. In Figure 9 we show the accuracy of the LR method as a function of the number of iteractions 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 Figure 9, all solvers have similar performance with the LBFGs slightly outperforming the other solvers for large number of iterations. In order to illustrate the probability domains in LR, we show the classification with two features (LBFGs, 200 iterations) in Figure 10. The domains look similar to the domains in the NN case (Figure 8). For the final classification we will use LBFGs solver with 200 iterations.

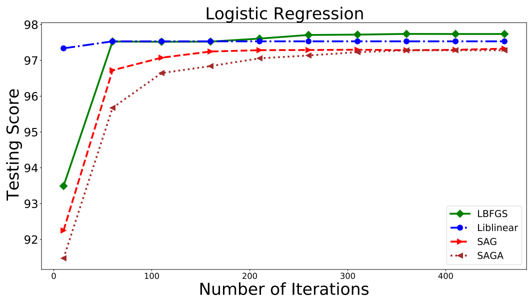

Fig. 9. Dependence of LR testing accuracy on the number of iterations for different solvers.

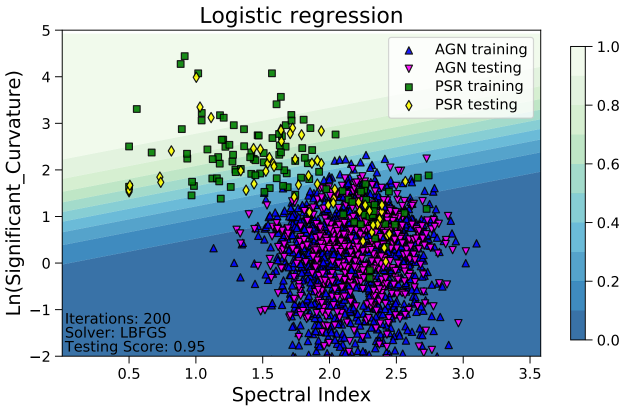

Fig. 10. Classification domains for LR with two features.

3.3. Oversampling

Fermi-LAT catalogs have many more AGNs than pulsars, i.e., the datasets are imbalanced. In the previous subsections we have optimized overall accuracy. 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 AGNs.

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 the pulsars, i.e., how complete is the list of predicted pulsars. 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.

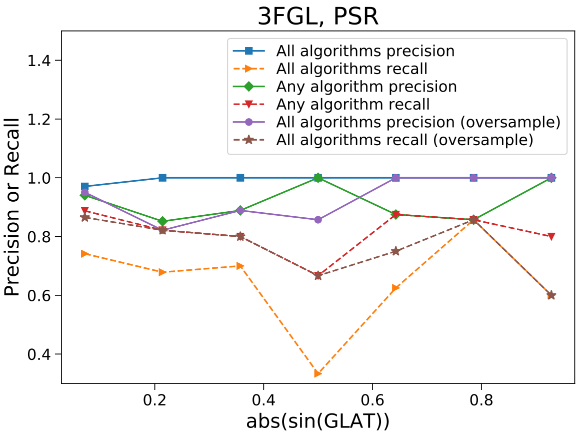

Fig. 11. Precision and recall for pulsars using all-algorithm and any-algorithm classification for unweighted training data and all-algorithm classification for pulsar oversampling in training. For details see Section 3.3.

In Figure 11 we show precision and recall for detection of pulsars using algorithms from previous sections. 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 as possible the list 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 have chosen oversampling. 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 Figure 12 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 Figure 12 we repeat Figure 10 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 the one without oversampling: 0.87 vs 0.95.

The results of training with oversampling are presented in Figure 11, 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 classify it as a pulsar. The recall in this case is similar to the recall of the any-algorithm-classification for the training without oversampling, while the precision is slightly better than the precision in the any-algorithm-classification (the precision is 1 for $|\sin(b)| > 0.6$).

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

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

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

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 10 neurons, Adam solver, and 300 epochs, and LR with LBFGS solver and 200 iterations. For training we use the pulsars and AGNs from the 3FGL catalog. 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.

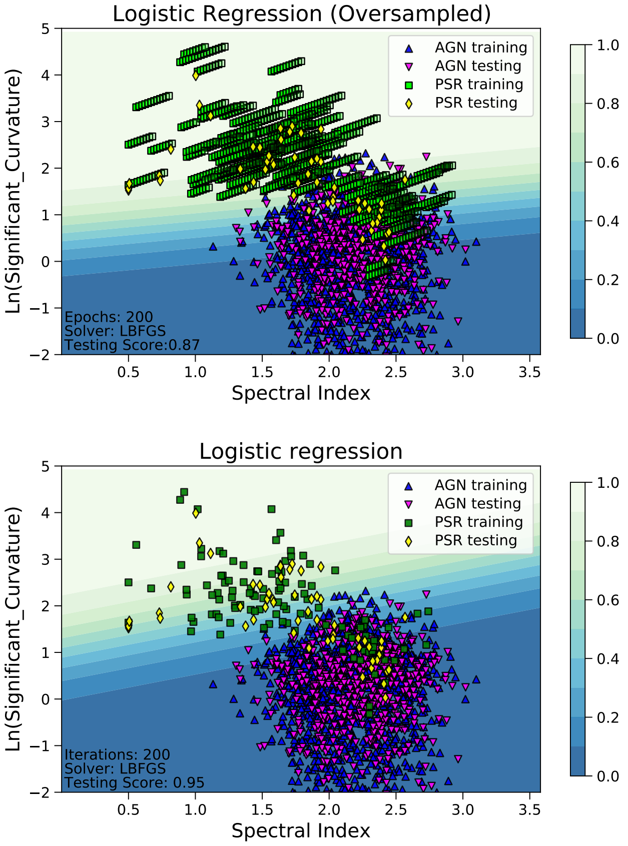

Fig. 12. Top panel: LR classification domains showing class probabilities for training with oversampling. The oversampling is illustrated by repeating the pulsar markers with a shift: the number of markers is equal to the number of times the pulsar appears in training. Bottom panel: we repeat Figure 10 for convenience of comparison with the oversampled training in the top panel.

Algorithm	Parameters	Testing Accuracy	Std. Dev.	Comparison with 4FGL Accuracy
RF	50 trees, max depth 6	97.42	0.60	96.1
RF_O		97.71	0.56	95.1
BDT	100 trees, max depth 2	97.80	0.55	95.34
BDT_O		97.82	0.51	93.8
NN	300 epochs, 10 neurons, Adam	97.40	0.87	94.55
NN_O		95.96	0.84	91.43
LR	200 iterations, LBFGS solver	97.60	0.56	93.48
LR_O		94.01	0.97	87.07

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

The selected algorithms are summarized in Table 2, where oversampling is shown by "_O". "Average testing accuracy" is computed by taking 1000 times 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: there are 278 such sources. The accuracy of our prediction

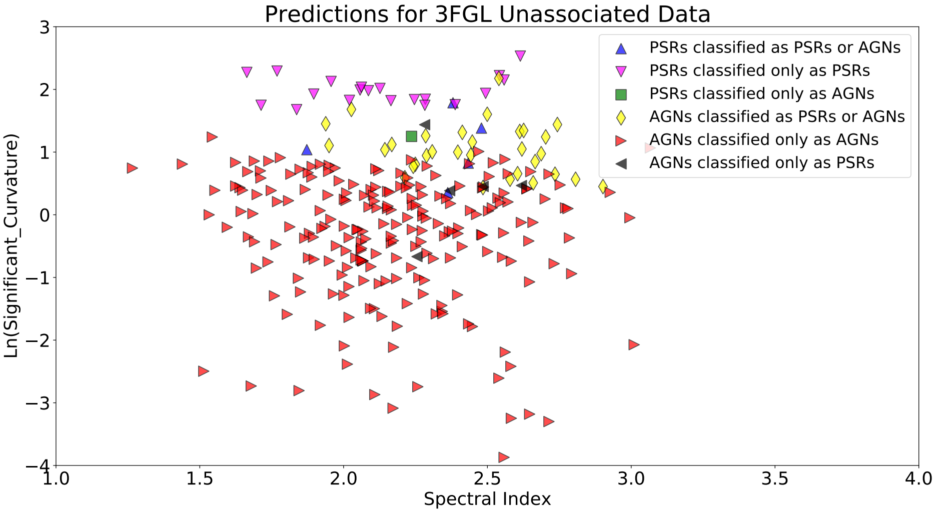

Fig. 13. Comparison of class prediction for unassociated 3FGL sources with classes in 4FGL. For more details see Section 4.1.

for the four selected algorithms with and without oversampling, taking the 4FGL classes as the true values, is reported in the column "Comparison with 4FGL Accuracy". The correct classifications and misclassifications for the 278 sources with associations in 4FGL are also presented in Figure 13. The class at the beginning of the label name corresponds to the association in the 4FGL, 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 PSR association in 4FGL 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 but the corresponding unassociated sources in 3FGL have both PSR and AGN classifications by different ML methods. The unassociated sources are classified as PSRs or AGNs if the corresponding probability is larger than 0.5. We notice that misclassified or partially misclassified sources in Figure 13 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.

As a result of the classification with the eight ML methods, we created a probabilistic catalog based on the 3FGL sources without missing values.[1] We train on 70% of the sources associated with pulsars or AGNs and save the probability values for testing sources, for sources which are not classified as pulsars or AGNs, 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 and sources not classified as AGNs or pulsars, while the average for AGNs and pulsar is over the number of times the sources appear in the testing sample, which is 300 on average.

We have also subselected the 278 unassociated 3FGL sources, which have PSR or AGN associations in 4FGL, and saved them for convenience of comparison as a separate file. 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 standard deviations of probabilities. Although class probabilities and standard deviation 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 possible generalizations to multi-class classification. Table 3 shows an example of the probabilistic catalog for a few unassociated 3FGL sources. Notice that the first source is classified as a pulsar by BDT and as an AGN by RF, LR, and NN algorithms, it is an example of a source with mixed classification. Out of 1008 unassociated sources in 3FGL, 96 are classified as pulsars by all eight methods, 580 are classified as AGNs, and 332 have mixed classifications. Out of 96 sources classified as pulsars, 6 sources have counterparts in Parkes survey (Camilo et al. 2015) within 2 arc minutes (see Table 4).

[1] There are thirteen sources with missing values in the 3FGL catalog (2 unassociated, 5 AGNs, 1 pulsar, and 5 "other" sources), which we save in a separate file "3FGL_sources_with_missing_values.csv" for reference.

	AGN Probability			
Source_Name_3FGL	BDT	RF	LR	NN
3FGL J1151.8-6108	0.046	0.671	0.753	0.626
3FGL J1417.5-4402	0.974	0.925	0.968	0.981
3FGL J1553.1+5437	0.999	0.983	0.999	0.995
3FGL J0002.6+6218	0.001	0.275	0.061	0.038

Table 3. Example of the AGN classification probabilities for a few unassociated sources in the 3FGL catalog (Acero et al. 2015).

Source_Name_3FGL	GLON	GLAT	Sep (arksec)
3FGL J0933.9-6232	282.2	−7.9	52.9
3FGL J1539.2-3324	338.7	17.5	60.5
3FGL J1744.1-7619	317.1	−22.5	46.1
3FGL J1753.6-4447	347.1	−9.4	18.2
3FGL J1808.3-3357	358.1	−6.7	103.7
3FGL J2112.5-3044	14.9	−42.4	88.3

Table 4. Connection of unassociated 3FGL sources classified as pulsars with Parkes pulsars (Camilo et al. 2015).

We summarize the results of classification of the unassociated 3FGL sources in Table 5. The “AGNs” column shows the number of unassociated sources where all eight methods from Table 2 give the probability for a source to be an AGN above 50%. Similarly the “Pulsars” column shows the number of unassociated sources where all four algorithms predict the source to be more likely 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. In the “Uncorrected” row we do not take into account that there can be sources other than AGNs or pulsars among the unassociated sources. We correct for the presence of the other sources by assuming 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, we denote by $N_{\rm AGN}$ the number of unassociated sources with AGN-like classification by all four algorithms, 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. The number of AGN-like sources among the unassociated source corrected for the presence of other sources is estimated as

$$N_{\rm AGN}^{\rm corr} = N_{\rm AGN} - N_{\rm AGN}^{\rm ass\,other} \frac{N_{\rm unass}}{N_{\rm ass}}. \qquad (2)$$

Analogous corrections are applied for the number of unassociated sources with pulsar-like classification by all eight methods, and for unassociated sources with mixed classification.

Correction for other sources	AGNs	Pulsars	Mixed
Uncorrected	580	96	332
Corrected	561.5	83.0	309.5

Table 5. Expected number of AGNs and pulsars among the unassociated 3FGL sources. “AGNs” and “Pulsars” columns show the number of sources where all eight methods predict the same class, “Mixed” column shows the number of sources with mixed classification. Uncorrected (corrected) rows show the number of predicted AGNs and pulsars among the unassociated sources, which are uncorrected (corrected) for the presence of sources other than AGNs and pulsars among the unassociated sources.

4.2. Probabilistic classification of sources in the 4FGL catalog

In this section we construct a probabilistic classification of sources in the 4FGL catalog (Abdollahi et al. 2020).[2] As in the previous section, we use for training and testing sources associated with either AGNs or pulsars, which have no missing values used for classification.[3] We then calculate the classification probabilities of AGN and PSR classes for both the associated and the unassociated sources. The 4FGL catalog has higher number of features, especially due to the difference in modeling of the spectra compared with the 3FGL catalog. We selected 31 of these features 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(Flux1000), PL_Index, Unc_LP_Index, LP_beta, LP_SigCurv, Unc_PLEC_Expfactor, PLEC_Exp_Index, hr12, hr23, hr34, hr45, hr56, hr67, ln(Variability_Index). Some of these features are directly related to the features, which we use in the 3FGL catalog, e.g., GLAT, PL_Index (instead of Spectral_Index), LP_SigCurv (instead of ln(Signif_Curve)), ln(Variability_Index), hardness ratios (in the 4FGL catalog there are two more energy bins compared to the 3FGL catalog).

For the classification of 4FGL sources, we do not perform another optimization of meta-parameters, i.e., 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. Similarly to the construction of the 3FGL probabilistic catalog, we

[2] There is also a version of the 4FGL catalog based on 10 years of data, rather than 8 years of data: the 4FGL-DR2 catalog (Ballet et al. 2020). We discuss our predictions for 4FGL-DR2 in Section 4.3.

[3] In the 4FGL catalog there is only one source with missing values: 4FGL J0534.5+2201i associated with the Crab pulsar wind nebula.

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 sources. The corresponding accuracies are reported in Table 6. All algorithms have a slightly better accuracy for the 4FGL catalog compared to the 3FGL catalog, which is likely due to a better determination of the spectra in 4FGL, to a higher number of features, and more associated sources used as training data.

Algorithm	Parameters	Testing Accuracy	Std. Dev.
RF	50 trees, max depth 6	98.27	0.35
RF_O		98.12	0.37
NN	300, 16 Neurons, Adam	98.08	0.39
NN_O		97.17	0.55
BDT	100 trees, max depth 2	98.23	0.40
BDT_O		98.15	0.36
LR	LBFGS solver, 200 iterations	98.08	0.35
LR_O		96.67	0.53

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

The expected numbers of AGNs and pulsars among the unassociated source in the 4FGL catalog are reported in Table 7. The definition of columns and rows is the same as in the 3FGL catalog case in Section 4.1.

Correction for other sources	AGNs	Pulsars	Mixed
Uncorrected	674	141	521
Corrected	638.5	119.1	475.5

Table 7. Expected numbers of pulsars and AGNs among unassociated sources in the 4FGL catalog (Abdollahi et al. 2020). For definitions see Table 5.

Finally, we looked at sources which were unassociated in both 3FGL and 4FGL. Out of 306 such sources, 38 sources are predicted to be pulsars using 3FGL features and 71 sources are predicted to be pulsars using 4FGL features. This leads to 29 sources which were predicted by all eight methods to be pulsars for features taken from both 3FGL and 4FGL catalogs. Among these 29 sources, four can be spatially associated to pulsars in the Parkes survey (of the other two, one is now associated as a pulsar in 4FGL, and the second one is not detected in 4FGL). For convenience, we save these 29 pulsar candidates as a separate file.

4.3. 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). There are 14 sources in 4FGL-DR2 with missing values: four AGNs, one PWN (Crab), and nine unassociated sources. The expected numbers of pulsars and AGNs among the 1670 unassociated sources in 4FGL-DR2 without missing values are presented in Table 8. Correction for the presence of other sources is calculated similarly to the 3FGL calculation in Section 4.1. We find that compared to the 4FGL catalog, there are more sources classified as AGNs and sources with mixed classification among the unassociated source in 4FGL-DR2, whereas the numbers of sources classified as pulsars in 4FGL and 4FGL-DR2 are similar.

Correction for other sources	AGNs	Pulsars	Mixed
Uncorrected	854	144	672
Corrected	801.9	120.8	602.0

Table 8. Expected numbers of pulsars and AGNs among unassociated sources in the 4FGL-DR2 catalog (Ballet et al. 2020). For definitions see Table 5.

5. Application of probabilistic catalogs for population studies

5.1. Number of sources as a function of flux

In this section we show how probabilistic catalogs can be used, for instance, for population studies. One of the most important questions in gamma-ray astronomy is contribution of point sources, e.g., AGNs, to the extra-galactic 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 extra-galactic 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 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

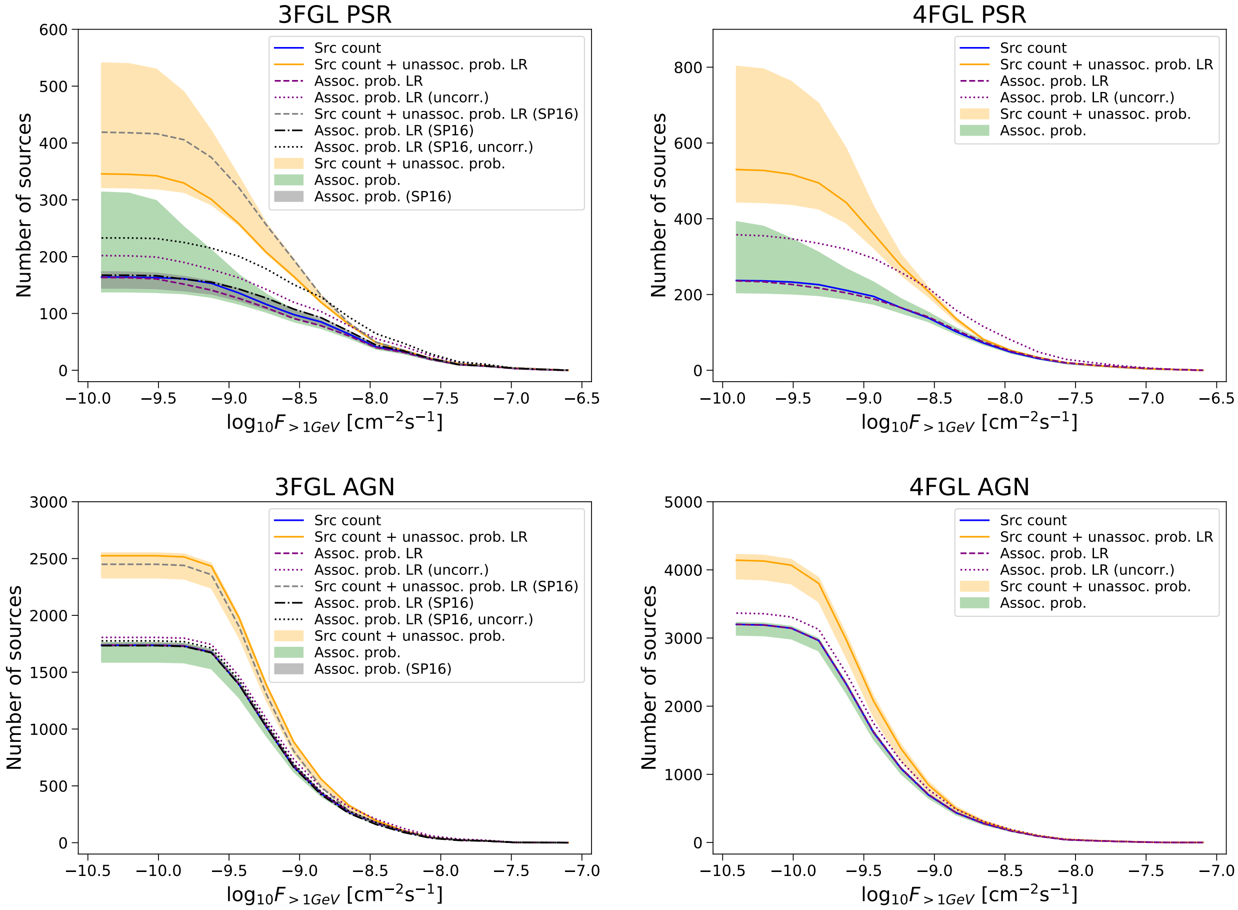

Fig. 14. Cumulative number of sources as a function of their flux. Green bands show the envelope of the sum of class probabilities for associated sources, while orange bands show the sum of counts of associated sources (blue solid line) plus the sum of probabilities for unassociated sources. The curves with "SP16" in the labels are derived from the data in Saz Parkinson et al. (2016). For details see Section 5.1.

section we determine the numbers of AGNs and pulsars as a function of their flux.

In Figure 14 we show the cumulative number of AGNs and pulsars with 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 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:

$$N^{\rm ass}_{\rm AGN} = \sum_{i\in {\rm ass}} p^i_{\rm AGN} - \sum_{i\in {\rm ass\,other}} p^i_{\rm AGN}. \tag{3}$$

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 of ML algorithms. It is important to note that 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 pulsars overpredicts the pulsar counts in 3FGL, while correction for "other" sources makes the prediction consistent with the counts of pulsars.

The predictions for the number of AGNs and pulsars among the unassociated sources corrected for "other"

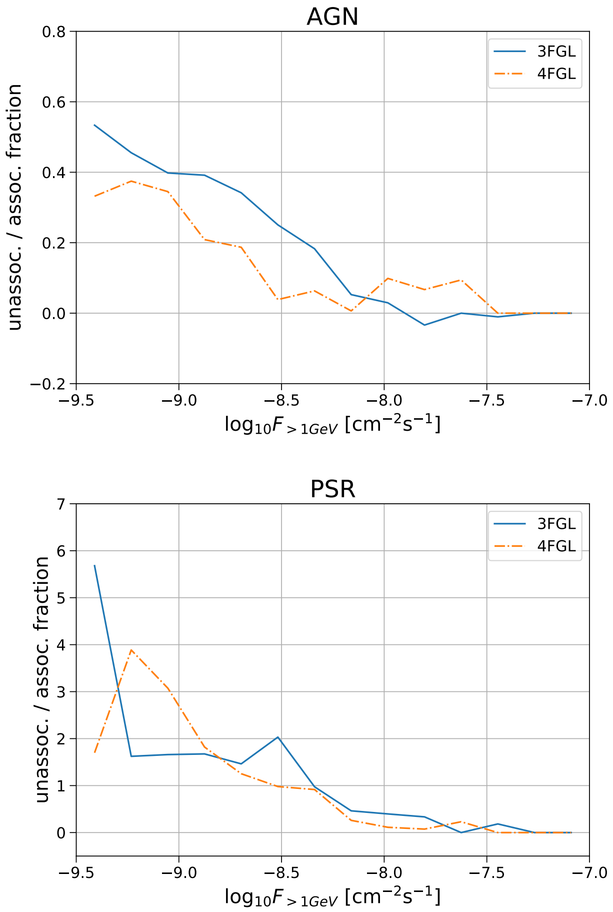

Fig. 15. Ratio of estimated number of AGNs and pulsars among unassociated sources corrected for the presence of other sources (Equation (4)) to the counts of associated AGNs and pulsars respectively.

sources added to the 3FGL and 4FGL source counts are shown by solid orange lines (for the LR case). 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 Equation 3, 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

$$N_{\rm AGN}^{\rm unass} = \sum_{i\in {\rm unass}} p_{\rm AGN}^i - \sum_{i\in {\rm ass\,other}} p_{\rm AGN}^i \cdot \frac{N_{\rm unass}}{N_{\rm ass}} \qquad (4)$$

where all probabilities and the numbers 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 LR method of Saz Parkinson et al. (2016) and corrected for other sources are shown by dashed grey lines.

We predict that the expected number of pulsars among the unassociated sources in the 3FGL catalog is 267 ± 110, where the range is the envelop of the sums of probabilities in Equation (4) for different ML methods (including oversampling) corrected for other sources among the unassociated sources. The expected number of pulsars among the unassociated sources in the 4FGL catalog corrected for other sources is 386 ± 179. These numbers are larger than the number of associated PSRs without missing values (164 in 3FGL and 237 in 4FGL). Even at the lower range of expected numbers of pulsars among unassociated sources, there are potentially as many pulsars as there are associated ones.

We note that according to Table 5, the number of unassociated 3FGL sources with $p_{\rm PSR} > 0.5$ for all four ML algorithms is 96 (83), while there are 332 (309.5) sources with mixed classification, uncorrected (corrected) for other sources. The number of sources with mixed classification (309.5 for 3FGL or 475.5 for 4FGL) is larger than the range of values for the expected number of pulsars calculated for the sum of probabilities (220 for 3FGL or 358 for 4FGL). It means that the decision which sources are considered to be more likely pulsars is more sensitive to the choice of the ML method and the probability threshold than the expected number of pulsars calculated from the sum of probabilities.

We also note that the probabilistic classification mostly affects sources with smaller fluxes, which we illustrate in Figure 15, where we show that the ratio of expected number of AGNs and pulsars among unassociated sources computed according to Eq. 4 using LR method without oversampling to the number of associated sources decreases as the flux increases. Negative values (e.g., at high fluxes for AGNs) are due to subtraction of probabilities for the "other" associated sources.

5.2. Latitude and longitude profiles

In this section we show Galactic latitude and longitude profiles of the distributions of associated and unassociated sources. In Figure 16 we present the source counts as a function of abs(sin(GLAT)), where 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 catalogs. It is interesting to note that the den-

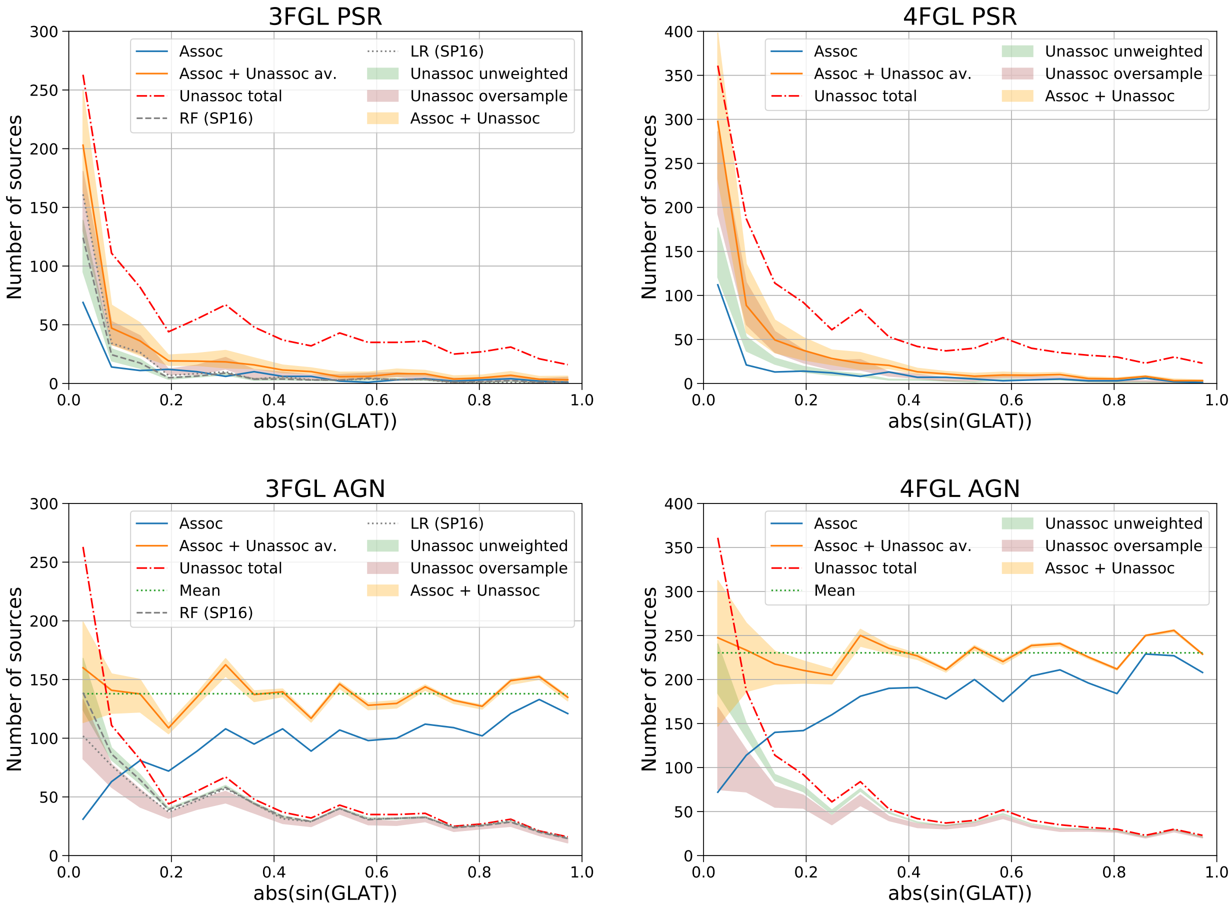

Fig. 16. Latitude profiles of source counts. Blue solid line – associated 3FGL and 4FGL sources. Red dash-dotted line – all unassociated sources. Green (red) band – envelope of sums of class probabilities for unassociated sources for the four ML algorithms without (with) oversampling. Orange solid line (band) – average (envelope) of sums of class probabilities for the eight ML methods with and without oversampling added to the source count of associated sources. Green dashed line on the AGN plots – mean of the sum of counts of associated sources and the average of the expectations for counts of unassociated sources (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 5.2.

sity of associated AGNs is decreasing near the Galactic plane. The total counts of unassociated sources are shown by red dash-dotted lines. Green shaded areas show the envelopes of sums of probabilities for AGN- and PSR-like sources for the four algorithms without oversampling, while the red shaded areas show the envelope for the four algorithms with oversampling. The classifications of 3FGL sources by Saz Parkinson et al. (2016) are shown by gray dashed (RF) and dotted (LR) lines. In this section we do not perform a correction for the presence of other sources among the unassociated ones. The numbers of unassociated sources classified as AGNs and PSRs grow towards the Galactic plane (GP). Within $\approx 4\overset{\circ}{.}5$ from the GP the expected number of PSRs is about the same as the number of AGNs among unassociated sources (the first data point on the left). At high latitudes, most of 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 algorithm. It can be a posteriori explained by the fact that the density of AGNs is such 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, but the sum of the two is relatively uniform as a function of Galactic latitude.

In Figure 17 we show plots analogous to Figure 16 for Galactic longitudes. We note that there is a significant increase in the number of unassociated sources in the 4FGL catalog for $|\ell| \lesssim 50°$. It leads to a large ex-

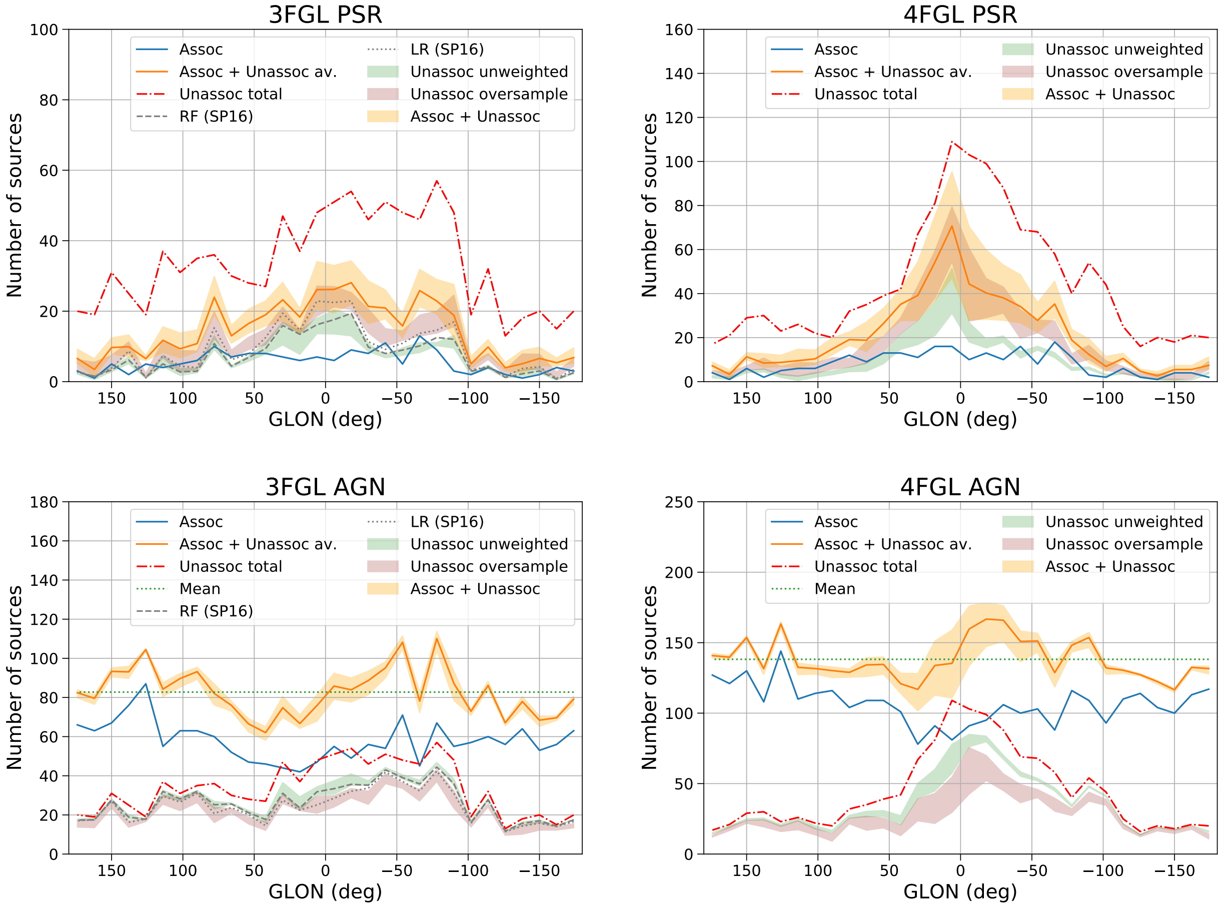

Fig. 17. Longitude profiles of source counts. For the definition of labels see Figure 16.

pected number of pulsars among unassociated sources for these longitudes. The number of associated AGNs is smaller than average for $|\ell| \lesssim 50°$, while the expected number of AGNs among unassociated sources is larger than average for these longitudes. The sum of the two is relatively uniform, with a possible overprediction of AGNs in the unassociated sources in the 4FGL catalog for $-50° < \ell < 0°$.

6. Conclusions

In the paper we determine the probabilities of classification of unassociated sources in the 3FGL, 4FGL, and 4FGL-DR2 *Fermi*-LAT catalogs. 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 have scanned some meta-parameters of the algorithms, such as depth of the trees, the number of trees, the number of neurons, in order to determine optimal parameters which do not create overfitting of data and provide good accuracy of classification. The accuracies, which we obtained for the 3FGL catalog for the four algorithms without oversampling, are about 97%, with oversampling the accuracies are between 94% and 97%. We have also checked the accuracy of classification by selecting unassociated sources in 3FGL, which have associations in 4FGL. If we take the 4FGL associations as the true value, then the accuracies of classification in this subset of sources with (without) oversampling are between 93% and 97% (87% and 95%). As one can see from Figure 13, the misclassified sources have spectral parameters in 3FGL which are typical of the other class, i.e., the misclassification can be due to problems with reconstructing the spectrum of the sources.

We have created catalogs with probabilistic classifications of sources, where for each source and for each class (i.e., AGNs and pulsars in our case) we report class probabilities for each of the four algorithms trained on both unweighted and oversampled datasets. We also provide individual standard deviations for all classification probabilities by 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. 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 his or her needs. For example, in a search of 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 class, in addition, one can use weighting by probability.

As an example of the application of the probabilistic catalogs, we derive the expected number of sources in the catalog as a function of their flux, including the unassociated sources. As a consistency check, we compare the counts of associated sources to the sums of probabilities for associated sources. We find that correcting for the contribution of sources other than AGNs and pulsars plays an important role for estimation of the expected number of sources in a particular class. We find the total expected number of AGNs and pulsars in 3FGL and 4FGL catalogs by adding the class probabilities for the unassociated sources to the source counts of associated sources and correcting 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 plot the counts of associated sources and the expected number of AGNs and pulsars among unassociated sources as functions 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, but the sum of the two is relatively uniform, as expected for extragalactic sources. We also find that the expected number of pulsars among unassociated 4FGL sources is significantly larger than average for longitudes $|\ell| \lesssim 50^\circ$.

Acknowledgements

The authors would like to thank Pablo Saz Parkinson for valuable discussions. This work 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).

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

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

Carr, B. J., Kohri, K., Sendouda, Y., & Yokoyama, J. 2010, Phys. Rev. D, 81, 104019

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

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

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

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

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

Nolan, P. L., Abdo, A. A., Ackermann, M., et al. 2012, ApJS, 199, 31

Robitaille, T. P., Tollerud, E. J., Greenfield, P., et al. 2013, A&A, 558, A33

Salvetti, D., Chiaro, G., La Mura, G., & Thompson, D. J. 2017, MNRAS, 470, 1291

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

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

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., Kang, S.-J., & Zheng, Y.-G. 2020, arXiv e-prints, arXiv:2001.06010

Appendix A: Tests of additional meta-parameters

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

LR algorithm has two hyper-parameters regularization and tolerance. As can be seen in figure A.1 the effect on accuracy is less than 1%. Therefore we used the default values for these parameters (tolerance is $1e^{-4}$ and regularization parameter is set at 1).

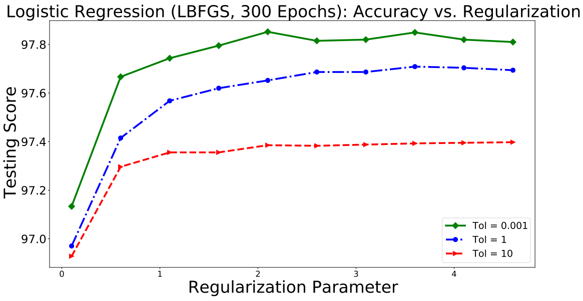

Fig. A.1. Dependence of LR on tolerance and regularization

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

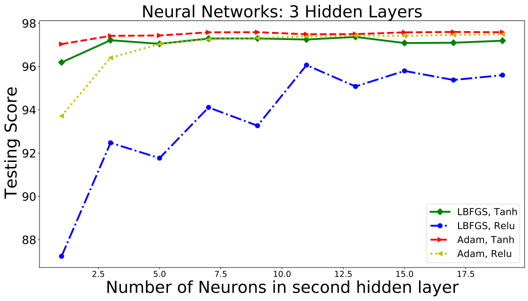

Fig. A.2. Dependence of NN on the second hidden layer, for 10 neurons in the first layer.

At the end we summarize features and their statistics, which we use for probabilistic classification of sources in the 3FGL and 4FGL catalogs, in Tables A.1 and A.2 respectively.

Feature Name	Mean	Standard Deviation	Minimum	Maximum
ln(Flux_DEN)	−28.21	2.51	−35.4	−19.88
ln(unc_energy_flux_100)	−27.47	0.47	−28.47	−24.78
Spectral_Index_3fgl	2.14	0.37	0.5	3.1
ln(Signif_curve)	0.23	1.19	−5.81	4.44
ln(var)	4.35	0.95	3	11.01
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
GLAT	2.67	41.1	−87.66	86.37

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

Feature Name	Mean	Standard Deviation	Minimum	Maximum
GLON	179.19	101.55	0.09	359.99
GLAT	2.04	40.96	−87.68	87.57
ln(Flux1000)	−21.55	1.2	−25.45	−13.51
PL_Index	2.22	0.3	1.4	3.49
Unc_LP_Index	0.16	0.13	0	2.93
LP_beta	0.15	0.19	−0.17	1
LP_SigCurv	2.77	6.66	0	207.66
hr12	0.06	0.74	−1	1
hr23	−0.27	0.61	−1	1
Unc_PLEC_Expfactor	0	0	0	0.05
hr34	−0.52	0.36	−1	1
hr45	−0.53	0.26	−1	1
hr56	−0.65	0.28	−1	1
hr67	−0.56	0.52	−1	1
ln(Variability_Index)	2.89	1.44	−1.08	10.84
ln(Pivot_Energy)	7.45	0.78	4.93	10.13

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