Effect of covariate shift on multi-class classification of *Fermi*-LAT sources

Dmitry V. Malyshev [ID]★

Erlangen Centre for Astroparticle Physics, Nikolaus-Fiebiger-Str 2, Erlangen D-91058, Germany

ABSTRACT

Probabilistic classification of unassociated *Fermi*-LAT sources using machine learning methods has an implicit assumption that the distributions of associated and unassociated sources are the same as a function of source parameters, which is not the case for the *Fermi*-LAT catalogues. The problem of different distributions of training and testing (or target) data sets as a function of input features (covariates) is known as the covariate shift. In this paper, we, for the first time, quantitatively estimate the effect of the covariate shift on the multi-class classification of *Fermi*-LAT sources. We introduce sample weights proportional to the ratio of unassociated to associated source probability density functions so that associated sources in areas, which are densely populated with unassociated sources, have more weight than the sources in areas with few unassociated sources. We find that the covariate shift has relatively little effect on the predicted probabilities, i.e. the training can be performed either with weighted or with unweighted samples, which is generally expected for the covariate shift problems. The main effect of the covariate shift is on the estimated performance of the classification. Depending on the class, the covariate shift can lead up to 10–20 per cent reduction in precision and recall compared with the estimates, where the covariate shift is not taken into account.

Key words: Machine Learning – Data Methods – Gamma Rays – Catalogues.

1. INTRODUCTION

Classification of unassociated *Fermi*-LAT sources with machine learning (ML) provides an opportunity to probabilistically determine the classes of sources based on their gamma-ray properties, when direct multi-wavelength association is not known (Ackermann et al. 2012; Mirabal et al. 2016; Saz Parkinson et al. 2016; Lefaucheur & Pita 2017; Luo et al. 2020; Finke, Krämer & Manconi 2021; Zhu, Kang & Zheng 2021; Bhat & Malyshev 2022; Malyshev & Bhat 2023). For some of the unassociated sources it may even be impossible to detect an associated source, e.g. for pulsars with a radio jet that is not pointing at the observer. In this case, the probabilistic classification of unassociated sources is the only possibility to determine the likely nature of the unassociated sources and to perform population studies including the unassociated sources.

One of the caveats of ML classification of *Fermi*-LAT sources is that the distributions of associated and unassociated sources in the feature space are different. For example, the fraction of associated sources at high latitudes is about 90 per cent, while the association fraction along the Galactic plane is about 50 per cent (Abdollahi et al. 2020). One of the reasons is that the density of gamma-ray sources is larger along the Galactic plane (GP), while there is also absorption in optical and soft X-ray bands by dust and gas, respectively, which complicates the detection of possible multi-wavelength counterparts. In Fig. 1 (upper left panel), we show the probability distribution functions (PDFs) for associated ('Assoc' label) and unassociated ('Unas' label) sources as a function of the

sine of Galactic latitude (Glat).[1] Another example is the brightness of the sources: brighter sources are associated more often than the dimmer ones. The distributions of the source detection significance for associated and unassociated sources are shown in Fig. 1 (upper right panel). On the lower panels of Fig. 1, we show the variability index – the significance of temporal variability, which is often used for identification of counterparts based on correlated flares in blazars or pulsed emission in pulsars, and the log parabola of the beta coefficient (LP_beta) – the curvature of the log parabola fit of the spectrum (curved spectra are more typical for Galactic sources, such as pulsars). We see that, generally, associated and unassociated sources have different distributions as a function of features. Differences in the distribution of the training set (associated sources) and the target set (unassociated sources) may lead to biased predictions, such as the classes of unassociated sources, as well as wrong estimates of the classification performance (Luo et al. 2020; Finke et al. 2021; Zhu et al. 2021; Bhat & Malyshev 2022).

The basic assumption of the ML classification is that the joint distribution of the input features x and output features y are the same for the training and target data sets:

$$p_{\text{train}}(x, y) = p_{\text{target}}(x, y), \tag{1}$$

while, in general, a data set shift represents a situation when the training and target distributions are different $p_{\text{train}}(x, y) \neq p_{\text{target}}(x, y)$.[2] The joint distribution can be written as a product of conditional

---

[1] In the paper, we use the 4FGL-DR4 (Ballet et al. 2023) file version 'gll_psc_v32.fit'.

[2] Often the target data set is called the test data set. In this paper, we use the associated sources both for training and for testing. In order to avoid possible confusion, we use target data set to denote unassociated sources, which has
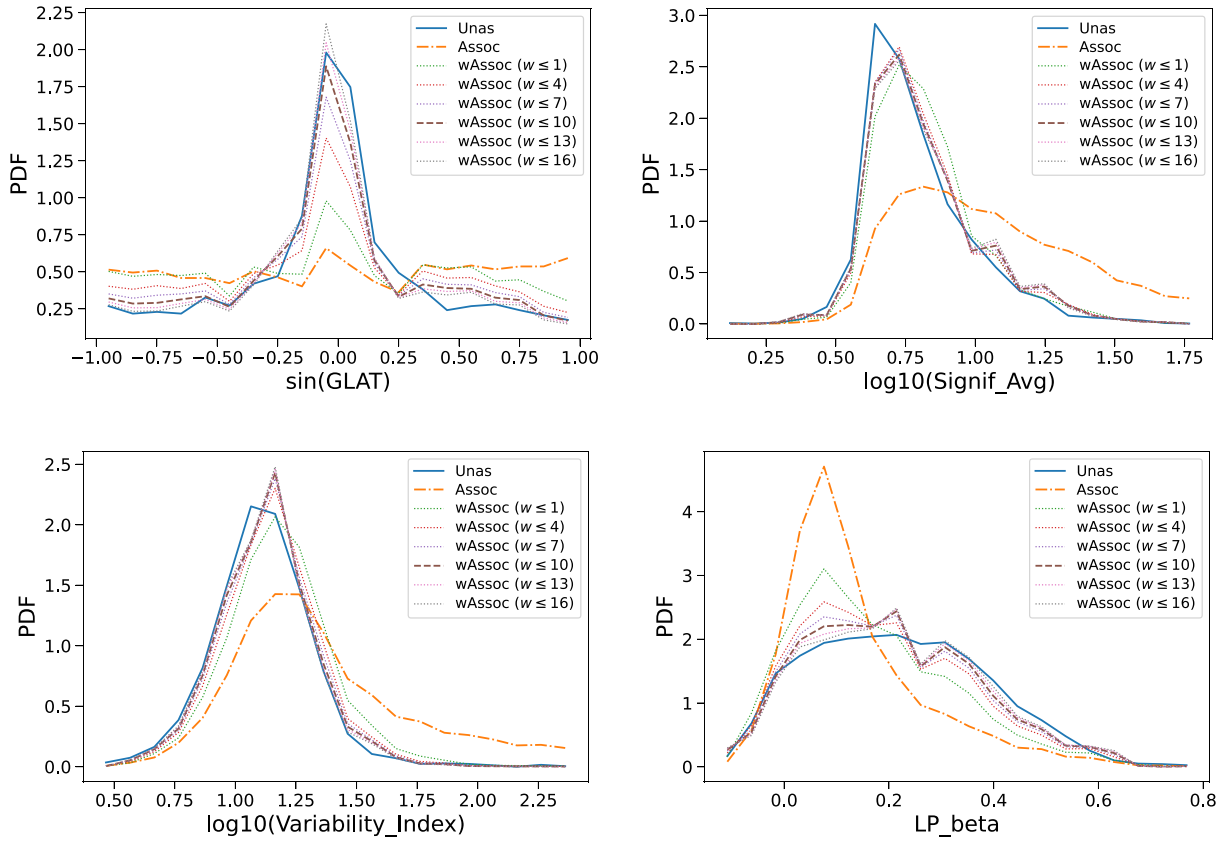
★ E-mail: dvmalyshev@gmail.com

Figure 1. PDFs of unassociated ('Unas') and associated ('Assoc') sources as a function of the sine of Glat (upper left panel), log of the average significance of the source (upper right panel), log of the variability index (lower left panel), and the curvature of the log parabola spectrum (lower right panel) in the 4FGL-DR4 catalogue (Ballet et al. 2023). The weighted PDFs of associated sources ('wAssoc' labels) are obtained by multiplying the associated sources with a weighting factor proportional to the ratio of PDFs of unassociated to associated sources in Equation (3). The PDFs are modelled by Gaussian mixture models (see text for more details). The values in parentheses show the maximal sample weights.

probability times a prior distribution in two different ways:

$$p(x, y) = p(y|x)p(x) = p(x|y)p(y). \qquad (2)$$

Correspondingly, there are two special cases of the data set shift (Moreno-Torres et al. 2012):

(1) Covariate shift: $p_{\text{train}}(y|x) = p_{\text{target}}(y|x)$, but $p_{\text{train}}(x) \neq p_{\text{target}}(x)$. It represents the situation, when the conditional probability of a class given the input features is unchanged, but the distributions of samples as a function of input features are different for training and target data sets.

(2) Prior shift: $p_{\text{train}}(x|y) = p_{\text{target}}(x|y)$, but $p_{\text{train}}(y) \neq p_{\text{target}}(y)$. It represents the situation, when the prior probabilities for classes change (e.g. the overall fraction of sources is different for training and testing data sets), while the distribution of input variables for each class is unchanged.

In this paper, we assume that the observational limitations and association biases, which lead to the differences in the distributions of associated and unassociated sources affect all source classes in the same way, i.e. that the conditional probabilities remain the same as a function of input features: $p_{\text{train}}(y|x) = p_{\text{target}}(y|x)$. In this case the data set shift corresponds to the shift of covariates (input features): $p_{\text{train}}(x)$

a different distribution from both the training and testing data sets sampled from associated sources.

$\neq p_{\text{target}}(x)$. The main goal of the paper is to determine the effect of the covariate shift on the multi-class classification of unassociated *Fermi*-LAT sources.

An independent test of classification performance has been obtained before by cross-matching predictions for unassociated sources in an earlier catalogue, e.g. the Third *Fermi*-LAT catalogue (3FGL), with the associations in the newer Fourth *Fermi*-LAT catalogue (4FGL, Luo et al. 2020; Finke et al. 2021; Zhu et al. 2021; Bhat & Malyshev 2022). It has been observed that the performance with the cross-matching method is worse than the performance estimated from the testing data sets sampled from the associated sources and it was argued that this decrease in performance is due to the covariate shift (Luo et al. 2020; Finke et al. 2021; Zhu et al. 2021; Bhat & Malyshev 2022). Nevertheless, there are several issues with the cross-matching method, which we address in this paper:

(1) The sample of sources in the cross-matching data set is not representative of the total population of unassociated sources.

(2) The reduction of performance can be partially due to uncertainties in the reconstruction of the source parameters (e.g. Bhat & Malyshev 2022). However, such uncertainties do not lead to covariate shift: If the intrinsic distributions of associated and unassociated sources are the same, i.e. $p_{\text{train}}(x) = p_{\text{target}}(x)$ and the uncertainties depend only on features x, then the observed distributions of training and target data sets remain the same. The results of the cross-matching method depend both on the uncertainties on the features x and on the covariate shift. In this work we separate the two effects and

determine the influence of only the covariate shift on the classification performance.

(3) The cross-matching method cannot be used for the latest catalogue (there is no newer catalogue that can be used for cross-matching).

The paper is organized as follows. In Section 2, we introduce the data and define the classes. We describe the model for the covariate shift in Section 3. In Section 4, we determine the effect of the covariate shift on two- and six-class classification of the *Fermi*-LAT sources. We construct probabilistic catalogues of the *Fermi*-LAT sources including the effect of the covariate shift in Section 5. The conclusions are presented in Section 6. In Appendix A, we discuss the models for the distributions of associated and unassociated sources in the feature space. We discuss the selection of input features and their importance in Appendix B. In Appendix C, we provide details about the neural network (NN) classification, whereas in the main body of the paper we use the random forest (RF) algorithm.

2. DATA SELECTION AND DEFINITION OF CLASSES

As input, we use the parameters, which describe the main features of the gamma-ray sources, such as the position on the sky, energy spectrum, and temporal variability. In particular, we use following 10 features derived from the source parameters in the Fourth *Fermi*-LAT catalogue data release four (4FGL-DR4, Ballet et al. 2023) (see also description in Abdollahi et al. 2022, Malyshev & Bhat 2023, and Appendix B for details on the feature selection): sin(GLAT), cos(GLON), sin(GLON), $\log_{10}$(Energy_Flux100), $\log_{10}$(Unc_Energy_Flux100), $\log_{10}$(Signif_Avg), LP_beta, LP_SigCurv, $\log_{10}$(Variability_Index), and the index of the log parabola spectrum at 1 GeV. Although there are many more parameters in the 4FGL-DR4 catalogue, most of the parameters either describe the same quantity (such as the Galactic and equatorial coordinates of the sources) or are highly correlated (Bhat & Malyshev 2022). It has also been shown that, at least in case of the two-class classification, relatively few input features, e.g. five, can provide an optimal classification performance (Luo et al. 2020). In this work, we use the features similar to the features used in (Luo et al. 2020; Bhat & Malyshev 2022) as well as in the earlier works, e.g. (Saz Parkinson et al. 2016), with some modifications described in Appendix B. Four sources in the catalogue have missing features: 4FGL J0358.4−5446 (nova), 4FGL J0534.5+2201i (pulsar wind nebula, PWN), 4FGL J1820.8−2822 (nova), and 4FGL J2010.2−2523 (flat spectrum radio quasar, FSRQ). We exclude these four sources from the analysis in this paper.

We use the labels for the classes of the gamma-ray sources from the 4FGL-DR4 catalogue (Abdollahi et al. 2022; Ballet et al. 2023) and consider identified sources (upper-case class names in 4FGL-DR4) and associated sources (lower-case class names) on the same footing. The physical classes of sources are summarized in Table 1.

We consider sources with unknown nature of the multi-wavelength counterpart (labelled as 'unk' in the catalogue) as unassociated sources. Overall, we have 4614 associated, and 2577 unassociated sources. Note that the total number of sources is 7191, which is less than the number of sources 7195 in the 4FGL-DR4 catalogue (Ballet et al. 2023) by the four sources with missing input features.

Provided that some of the physical classes have too few members for a reasonable classification (e.g. less than 10 associated sources), we use a hierarchical procedure to determine the classes (Malyshev & Bhat 2023) that combines several physical classes with similar

Table 1. Classes of associated sources in the 4FGL-DR4 catalogue (Abdollahi et al. 2022; Ballet et al. 2023). Both associated and identified sources in the catalogue are referred as associated sources in this work.

Physical class	Associated sources	Description
gc	1	Galactic centre
psr	141	Young pulsar
msp	179	Millisecond pulsar
pwn	21	Pulsar wind nebula
snr	45	Supernova remnant
spp	124	Supernova remnant and/or pulsar wind nebula
glc	34	Globular cluster
sfr	6	Star-forming region
hmb	11	High-mass binary
lmb	9	Low-mass binary
bin	10	Binary
nov	6	Nova
bll	1490	BL Lac type of blazar
fsrq	819	FSRQ type of blazar
rdg	53	Radio galaxy
agn	8	Non-blazar active galaxy
ssrq	2	Steep spectrum radio quasar
css	6	Compact steep spectrum radio source
bcu	1624	Blazar candidate of uncertain type
nlsy1	8	Narrow-line Seyfert 1 galaxy
sey	3	Seyfert galaxy
sbg	8	Starburst galaxy
gal	6	Normal galaxy (or part)

distributions in the feature space. In particular, we use the Gaussian mixture model (GMM) to determine the hierarchical splitting of the physical classes (for details, see Malyshev & Bhat 2023). An example of such splitting of the physical classes with the condition on the minimal number of sources in a class $n_s > 50$ is shown in Fig. 2, top panel. We note that node 011, which includes rdg, sey, sbg, and agn classes, has only 74 associated sources. We have checked that the classification into seven classes corresponding to the terminal nodes in the top panel of Fig. 2 does not give reasonable results for the 011 class. However, if we increase the condition on the minimal number of sources to be, e.g. $n_s > 100$, then node 01 cannot split. As a result, this node has 3194 sources, which is almost 70 per cent of all associated sources. In this case, the multi-class classification is not meaningful either. A possible solution to this problem is to first construct the classes with the condition $n_s > 50$ and then prune the tree by removing the node with the minimal number of classes, e.g. node 011. The removed node is shown by the dashed box in the top panel of Fig. 2. Since the parent node 01 has now only one child node 010, we merge nodes 010 and 01, i.e. the subtree under 010 is now a subtree under 01 (the corresponding nodes in the subtree move one level up). The resulting tree is shown in Fig. 2, bottom panel.

The physical classes in the pruned node are distributed among the remaining six classes. In order to determine that, we train RF classification with the six classes and then classify sources in rdg, sey, sbg, and agn classes using the six-class classification.[3] We compute the sum of class probabilities for all sources in each of the rdg, sey,

---

[3]In this paper, we use RF with maximal number of trees 50 and maximal depth of 15. For the other parameters we use default values in the SCIKIT-LEARN version 1.2.2 (Pedregosa et al. 2011) implementation of RF. In particular, the Gini index is used for the determination of the splits.
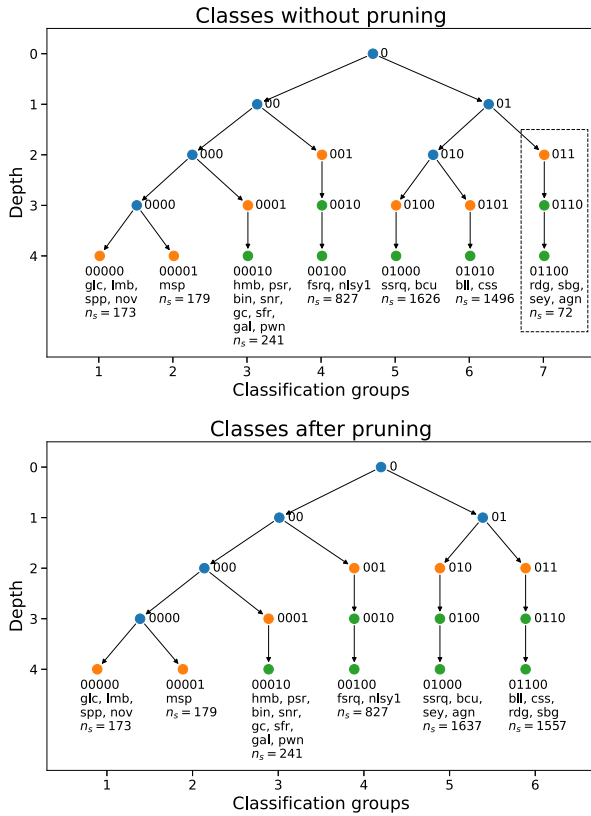
Figure 2. Definition of classes. Top panel: Hierarchical definition of classes (following the method of Malyshev & Bhat 2023) with the condition that the number of sources in a class is larger than 50. Dashed box shows the class with the smallest number of sources. We remove this class from the final definition of classes. The corresponding physical classes (rdg, sey, sbg, and agn) are distributed among the remaining six classes according to the maximal class probability sum for each of the physical class. Since node 01 has only one child 010 after pruning, we merge the nodes 01 and 010 into a new node 01. Bottom panel: The result of pruning, which shows the final hierarchical structure of the classes used for the classification in this paper. See text for more details.

Table 2. Definition of classes. The classes are labelled by the largest physical class, e.g. spp+ or msp+.

Class label	Physical classes	Associated sources
spp+	glc, lmb, spp, nov	173
msp+	msp	179
psr+	hmb, psr, bin, snr, gc, sfr, gal, pwn	241
fsrq+	nlsy1, fsrq	827
bcu+	ssrq, bcu, sey, agn	1637
bll+	bll, css, rdg, sbg	1557

sbg, and agn classes and attach these classes to groups with the largest sum of class probabilities. The result is that the sey and agn classes are attributed to node 010 dominated by the bcu class, while the rdg and sbg classes are attributed to node 011 dominated by the bll class. The result of this procedure is shown in the bottom panel of Fig. 2. We also show the summary of the remaining six classes in Table 2.

We note that physical classes are grouped in the six groups according to their gamma-ray properties, i.e. even if the physical nature of the sources is different but the gamma-ray properties

are similar, the physical classes would be added to the group. For example, the 'bll+' group has mostly active galactic nuclei (bll, css, and rdg) and the starburst galaxies class (sbg). A comparison of the two most important features for the separation of the physical classes at level one ('LP_beta' and 'log10(Unc_Energy_Flux100)') for bll, sbg, and several other physical classes can be found in fig. 1 of Malyshev & Bhat (2023). The assembly of the groups according to the similarities in the gamma-ray properties ensures an optimal multi-class classification performance. This is a fundamental limitation of any ML classification that the gamma-ray properties used for the classification do not necessarily reflect the different physical nature of the sources. In particular, in our analysis the ML classification cannot separate starburst galaxies from other sources in the bll+ class. Although the multi-class classification is dominated by the large classes, such as spp, msp, psr, fsrq, bcu, and bll, and, as a result, it is also mostly useful for the separation of these large classes, it nevertheless, can be useful also for the small classes, as it can provide additional evidence for association or classification. For example, if there is a nearby starburst galaxy in a vicinity of an unassociated source with high bll+ classification probability, then the source is more likely to be a starburst galaxy compared with the situation, when this source is classified as a member of, e.g. msp+ or fsrq+ classes.

3. COVARIATE SHIFT MODEL

The presence of the covariate shift manifests itself in the fact that the ratio of the training and the target PDFs is not a constant in the multi-dimensional feature space. Provided that the domains of the training and the target data sets are the same for the associated and unassociated sources, the effect of the covariate shift can be modelled by introducing weights for samples in the training and testing data set proportional to the ratio of the corresponding PDFs.

$$w(x_i) = \frac{p_{\text{unas}}(x_i)}{p_{\text{assoc}}(x_i)}. \tag{3}$$

In this case, the differences in the densities of training or testing and target data sets are compensated by the weighting of the samples. In order to determine the weighting factor $w(x)$ as a function of the features, one needs to model the PDFs $p_{\text{unas}}(x_i)$ and $p_{\text{assoc}}(x_i)$. There are different ways to approximate a distribution of discrete points with a continuous PDF, e.g. using kernel density estimators. In this paper, we use GMMs for modelling the PDFs of associated and unassociated sources. Details about the construction of the PDF models are provided in Appendix A. In order not to give too much weight to any of the sources, we put an upper bound on the weights. Examples of the PDFs for the associated sources including sample weights are presented in Fig. 1 ('wAssoc' labels) with several values of the maximal weight, e.g. $w \leq w_{\text{max}} = 1, 4, ..., 16$. Larger maximal weights typically give a better agreement between the distribution of unassociated sources and the weighted associated sources, especially for the Glat distribution. However, for most of the features the dependence on w_{max} is not very significant. Also larger maximal weight reduces the effective number of samples, where for a set of samples with weights w_i, the effective number of samples is computed as (Kish 1965):

$$n_{\text{eff}} = \frac{(\sum_i w_i)^2}{\sum_i w_i^2}. \tag{4}$$

We show the effective sample number as a function of w_{max} in the top panel of Fig. 3. For example, the effective number of associated sources for $w_{\text{max}} = 10$ is 961, which is more than four times
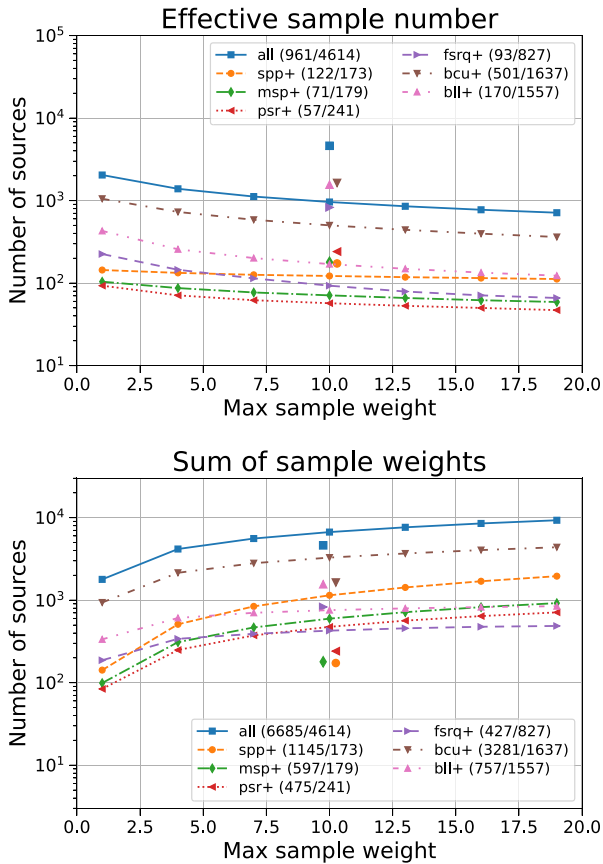
Figure 3. Effective number of samples (top panel) and oversampling (bottom panel) as a function of the maximal sample weight. Top panel: Effective number of samples as defined in Equation (4). The numbers in parentheses show, respectively, the effective number of samples at $w_{max} = 10$ and the total number of associated sources in each of the six classes (Table 2). The corresponding numbers of associated sources are also shown as the stand-alone points near $w_{max} = 10$. Bottom panel: Oversampling (or undersampling) of sources defined by summing the sample weights. The first number in parentheses shows the oversampled number of sources for all associated sources and for each of the six classes at $w_{max} = 10$. The second number is the number of associated sources (also shown as stand-alone points).

smaller than the total number of associated sources in the 4FGL-DR4 catalogue (4614), i.e. the variance of model parameters in the weighted sample case can be expected to be larger than in the unweighted case.

The weighting affects different classes unequally (see Table 2 for the class definitions). In particular, the effective number of fsrq+ sources is about nine times smaller than the number of associated fsrq+ sources, while the effective number of spp+ sources is less than 1.5 times smaller than the number of associated spp+ sources. Another characteristic number is oversampling (or undersampling), which is computed as the sum of the weights. The sum of weights for all sources and for the six classes is shown in the bottom panel of Fig. 3. We see that bll+ and fsrq+ classes are undersampled by a factor of about 2, while all other classes (including bcu+) are oversampled with an oversampling factor of up to 6.6 (for the spp+ class). Overall, we find that $w_{max} = 10$ provides a reasonable compromise between the approximation of the distribution of the unassociated sources (Fig. 1) and the effective number of samples (top panel of Fig. 3). We use the $w_{max} = 10$ case below for training and testing with weighted samples.

4. EFFECT OF COVARIATE SHIFT ON CLASSIFICATION

4.1 Two-class classification

In this section, in order to get an intuition about the effect of the covariate shift, we use a two-class (rather than a multi-class) classification problem. We define the two classes as active galactic nuclei (AGNs) and Galactic sources (including other galaxies):

AGNs: bll, fsrq, rdg, agn, ssrq, css, bcu, nlsy1, and sey (4013 sources);

Galactic: psr, msp, gc, pwn, snr, spp, glc, sfr, hmb, lmb, bin, nov, sbg, and gal (601 sources).

The classification is performed with the RF algorithm. A comparison of receiver operating characteristic (ROC) curves for the unweighted two-class classification and for the classifications where weights are applied both for training and testing is presented in Fig. 4. We see that the area under the curve (AUC) is reduced from about 0.96 to 0.89 both for AGNs and for Galactic sources.

The corresponding comparison of precision and recall is shown in Fig. 5. In this case AGNs are affected slightly more than the Galactic sources relative to statistical uncertainty.

It is interesting to note that if we use weighting for testing only and perform training with unweighted samples, then the performance is similar to the case when the weighting is applied both for training and for testing. We compare the corresponding precision and recall in Fig. 6. This result is not surprising, provided that the classification algorithm learns the conditional probabilities $p(y|x)$ that are not affected by the weights. Nevertheless, it does serve as a cross-check of the procedure, which shows that training with either weighted or unweighted samples can be used for the classification of unassociated sources. But it is important to use weights for the testing samples in the estimation of the performance to make sure that it is estimated for the sources with a distribution similar to the distribution of the target data set, i.e. the unassociated sources.

4.2 Multi-class classification

In this section, we study the effect of the covariate shift for training and testing in multi-class classification of the *Fermi*-LAT sources. For the classification, we use the six classes summarized in Table 2 and Fig. 2. As an example, we perform the classification with the RF algorithm in this section, while in Appendix C we compare the results with the classification using NN implemented with TensorFlow (Abadi et al. 2015). We use 70/30 per cent split into training and testing data sets. The performance is evaluated on the testing data sets.

The ROC curves calculated using the one-versus-all definition of the true positive and the false positive rates (TPR and FPR, respectively) are shown in the bottom panels in Fig. 7. Both the training and the testing are performed with weighted samples, where the weights are determined in equation (3). The shaded areas show the standard deviation of the ROC curves calculated from 10 random split into training and testing data sets. The second (from the bottom) row of panels in Fig. 7 shows the ROC curves in the five-class classification, where the five classes are obtained by merging some of the six classes by removing the last digit in the class names for the six classes (shown in the titles of the panels). In particular, the physical classes corresponding to the 0000 node are obtained by joining the classes in 00 000 and 00 001 nodes. The physical classes in the 0000 node are glc, lmb, spp, nov (which come from node 00 000), and msp (which come from node 00 001). The green curves show the ROC
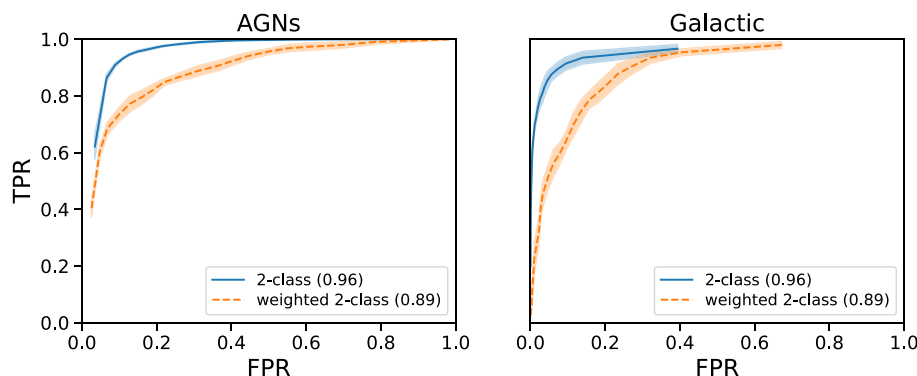
Figure 4. ROC curves for unweighted training and testing ('two-class' labels) and weighted training and testing ('weighted two-class' labels). The lines (shaded areas) represent the mean (the standard deviation) over ten splits into training and testing data sets with the 70/30 per cent ratio.
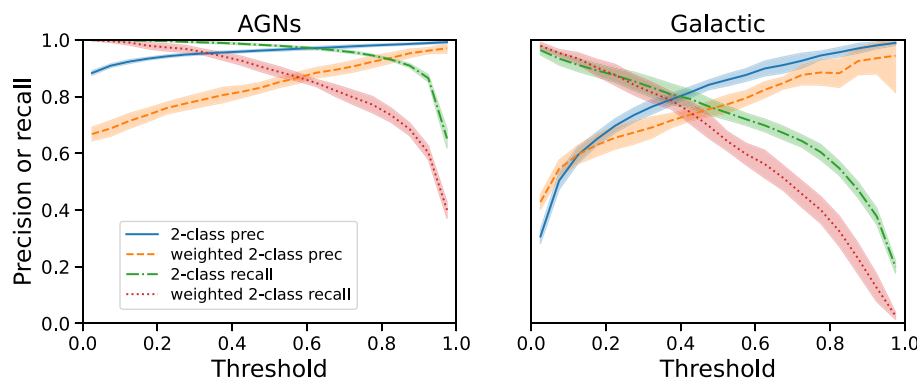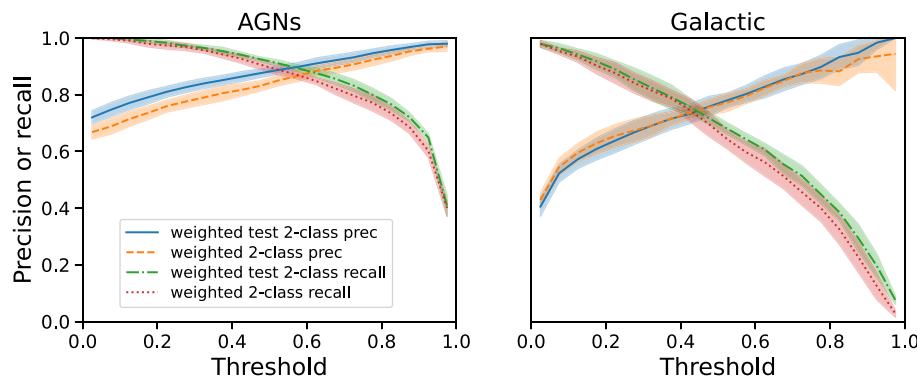

Figure 5. Comparison of precision and recall for unweighted training and testing ('two-class' labels) and weighted training and testing ('weighted two-class' labels). The lines (shaded areas) represent the mean (the standard deviation) over ten splits into training and testing data sets with the 70/30 per cent ratio.

Figure 6. Comparison of precision and recall for unweighted training with weighted testing ('weighted test two-class' labels) and weighted training and testing ('weighted two-class' labels). The performance calculated on weighted test samples for training with unweighted samples is similar to the performance of training with weighted samples. See Fig. 5 for the definition of the lines.

curves in the five-class classification. Analogously to the six-class classification, the green shaded area shows the uncertainty due to the random splits into training and testing data sets. The red curves in these panels show the ROC curves for probabilities obtained by summing the six-class probabilities of the children nodes. The performance for the direct five-class classification and for the five-class probabilities determined by summation of the six-class probabilities are practically the same. This conclusion also holds for the two- and four-class classification shown in rows one and two of Fig. 7, where we show both the ROC curves for the direct two- and four-class classifications and for probabilities determined by summation of

class probabilities of the children nodes. Similarly to the conclusions of Malyshev & Bhat (2023), we find that the performance of direct classification with two, four, and five classes and the performance of classification obtained by summation of probabilities of children nodes in the six-class case are very similar also for the weighted samples. Consequently, it is sufficient to do a classification with the maximal number of classes (six classes in this case).

In Fig. 8, we compare the ROC curves for the weighted and unweighted multi-class classifications. Similarly to the two-class case, the performance in the unweighted data set is better than for the weighted one. Provided that the unweighted data set represents the
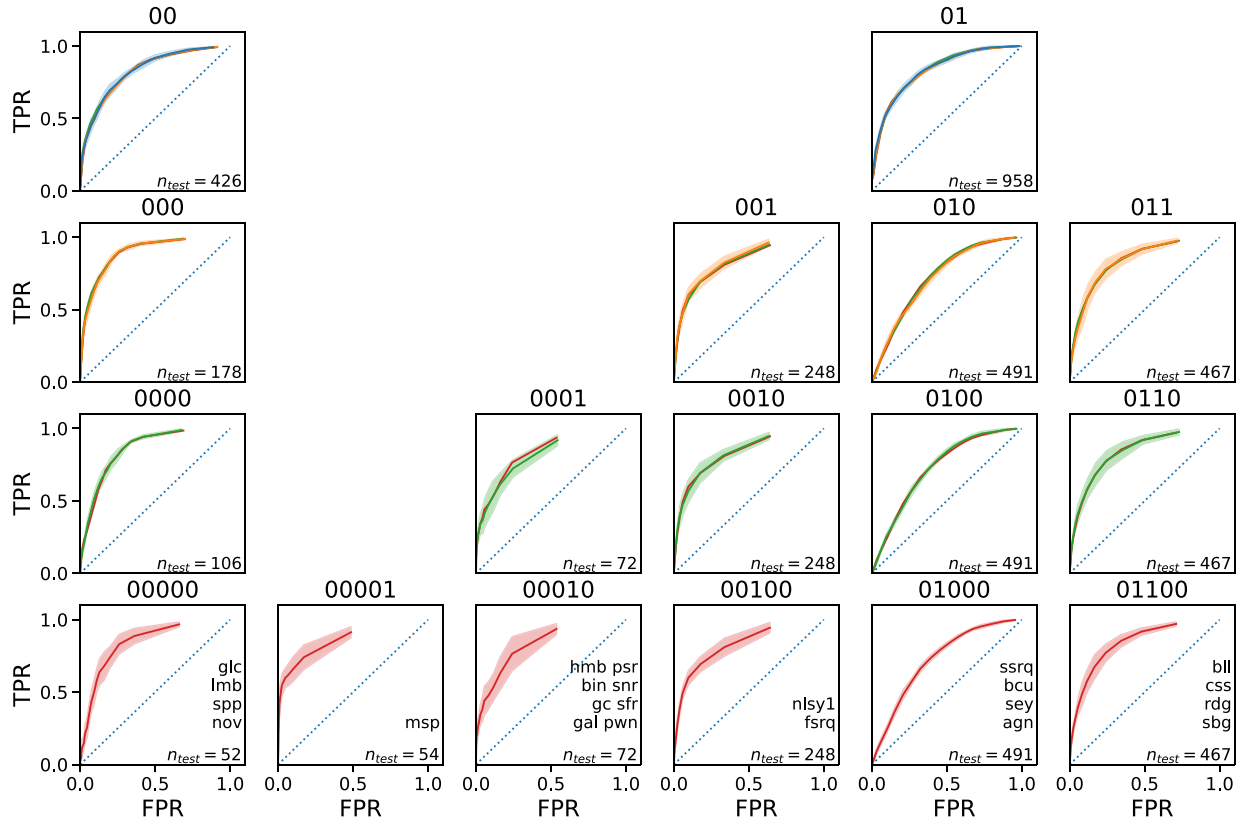
Figure 7. ROC curves for weighted training and testing following the hierarchical definition of classes in Fig. 2. The physical classes in a parent node are obtained by removing the last digits in the node names of the children nodes, e.g. the 0000 node contains physical classes of 00 000 and 00 001 nodes. At each level, the class probabilities are computed either directly for two-, four-, five-, or six-class classification or by summing the probabilities of the children nodes. Lines (shaded areas) show the mean (standard deviation) for 10 random splits into training and testing sets with 70/30 per cent ratio.
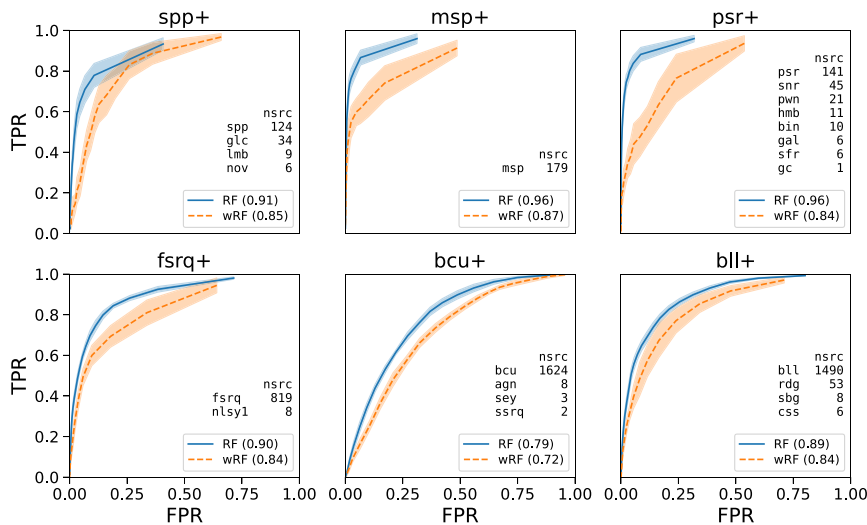

Figure 8. ROC curves for unweighted training and testing using the RF classification algorithm ('RF' labels) and weighted training and testing ('wRF' labels) for the six-class classification of the *Fermi*-LAT sources. The physical classes in each group and the numbers of associated sources in each physical class are shown in tables inside the panels.

associated sources, while the weighted one models the distribution of the unassociated sources, the performance of the classification for the unassociated sources determined from the associated sources (without weighting) is overestimated.

We show the difference in precision and recall for weighted and unweighted training and testing in Fig. 9. For example, the precision and recall for the bll+ and bcu+ classes in the weighted case are worse, respectively better, than in the unweighted case, while the
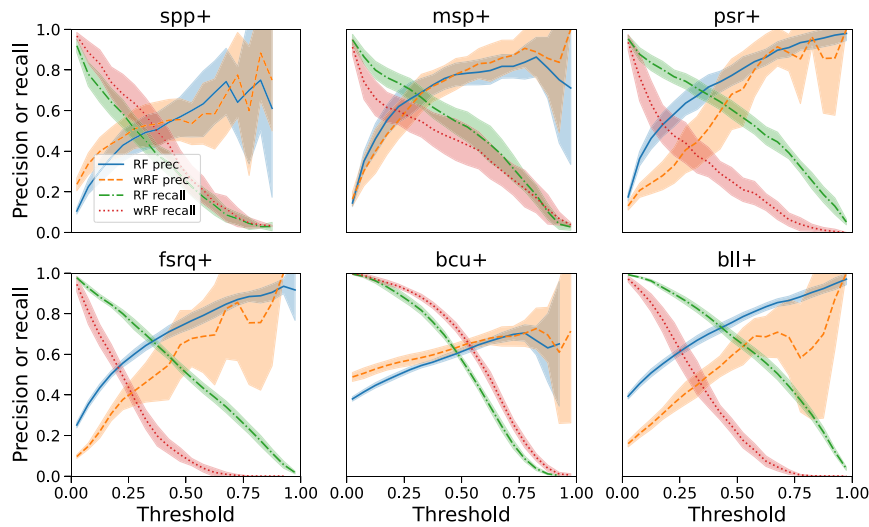
Figure 9. Precision and recall for unweighted training and testing using the RF classification algorithm ('RF' labels) and weighted training and testing ('wRF' labels) for the six-class classification of the *Fermi*-LAT sources. For the definition of the classes, see Fig. 8.

effect of using the weighted samples on the ROC curves for the bll+ and bcu+ classes are comparable, i.e. the AUC is smaller for both classes for the weighted relative to the unweighted cases. The worse ROC curves for the bll+ class is explained by the worse TPR, i.e. the recall, while the reduction of the ROC curve performance for the bcu+ class is explained by the worse FPR, which is the fraction of non-bcu+ sources attributed to the bcu+ class. We note that the bll+ class is undersampled by about a factor of two, while the bcu+ class is oversampled by almost a factor of two (bottom panel of Fig. 3). Another example of a slightly better precision and recall in the weighted case but a worse ROC curve is provided by the spp+ class. This can be explained by a similar increase in false positive and true positive detections but a smaller size of the 'negative' data set due to large oversampling of the spp+ class by a factor of 6.6 (bottom panel of Fig. 3).

In Figs 10 and 11, we compare the ROC curves and precision and recall for the classification trained on unweighted samples but tested on the weighted samples ('wtRF' labels) with the classification where both training and testing were performed on the weighted samples ('wRF' labels). We find a similar performance for weighted and unweighted training estimated from the tests on weighted samples, which is generally expected for the covariate shift. We also find that the probabilities are generally well calibrated both for weighted and unweighted training when tested with the weighted samples (Fig. 12).

5. CATALOGUE CONSTRUCTION WITH COVARIATE SHIFT

In this section, we describe the construction of probabilistic catalogues where the class probabilities are calculated using both weighted and unweighted training samples. As in the previous section, we use six classes and ten input features. For classification, we use RF and NN algorithms. In order to estimate the uncertainty of prediction due to the random choice of the training samples, we perform several 70/30 per cent splits into training and testing data sets. The predicted class probabilities for unassociated sources are computed as the mean over all predictions, while for the associated sources, the probabilities are determined by the mean over the splits where a source is included in the testing sample. In order to have a reasonable statistics for associated sources, we require that each

associated source appears in the testing data set at least five times, which resulted in 51 splits into training and testing data sets. In the catalogue, we report both the average class probabilities and the standard deviation of the class probabilities due to the random training/testing splits for each source. Thus, for each source we report six average class probabilities determined with the RF algorithm, six class probabilities determined with the NN algorithm, and the corresponding standard deviations (24 columns in total). We also include a column with sample weights. For the associated sources, the sample weight is equal to the ratio of the unassociated to associated sources PDFs with the maximal weight of 10. The sample weight for the unassociated sources is set to one.

We perform the classification using weighted and unweighted samples for training. The predicted numbers of associated and unassociated sources in the six classes (calculated as the sum of class probabilities) in the weighted and unweighted training cases are shown in Tables 3 and 4, respectively. The uncertainties are calculated as the root mean squared (RMS) of the corresponding standard deviations. It is interesting to note that the predicted number of sources in a class among unassociated sources is similar for the RF algorithm and for most of classes for the NN algorithm. However the expected number of sources in a class for associated sources is clearly biased in the weighted training case. Overall, we find that unweighted training provides a more reasonable result than the weighted training, because the performance evaluated on the weighted test samples is similar for weighted and unweighted training (cf. Figs 10, 11, and 12), while the predictions for the unweighted test samples are biased in the weighted training case (Table 3) and they are not biased in the unweighted training case (Table 4).

We compare the changes in the individual class probabilities for the unassociated sources for weighted and unweighted training in Figs 13 and 14. In Fig. 13, we calculate the difference of predicted class probabilities in four cases: weighted RF minus unweighted RF probability ('wRF − RF' label), weighted NN minus unweighted NN probability ('wNN − NN' label), unweighted RF minus unweighted NN probability ('RF − NN' label), and weighted RF minus weighted NN probability ('wRF − wNN' label). We see that in all cases the differences of the class probabilities for the individual sources is within about 13 per cent. The smallest differences (within about 4 per cent) are among weighted RF and unweighted RF probabilities
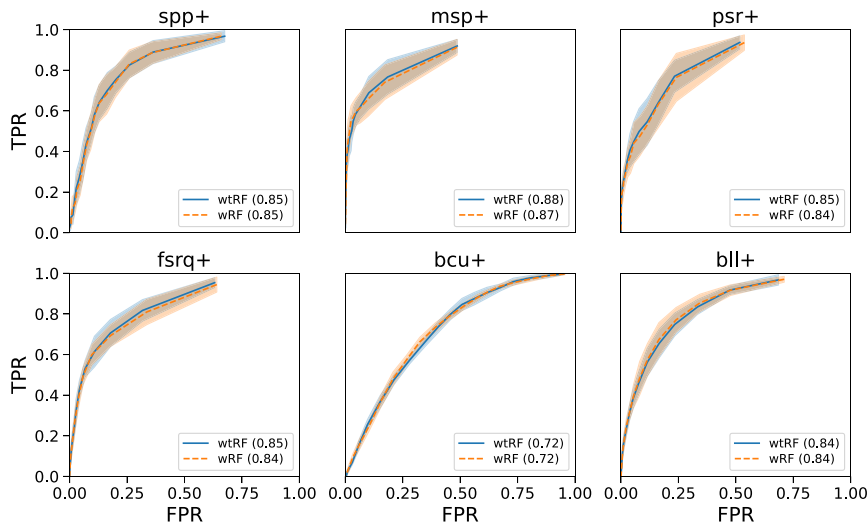
Figure 10. ROC curves for unweighted training but weighted testing using the RF classification algorithm ('wtRF' labels) and weighted training and testing ('wRF' labels) for the six-class classification of the *Fermi*-LAT sources.
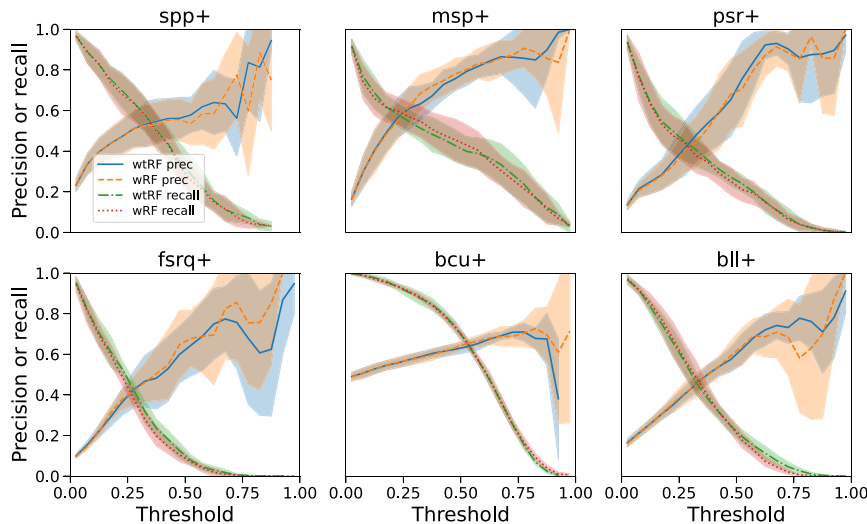

Figure 11. Precision and recall for unweighted training but weighted testing using the RF classification algorithm ('wtRF' labels) and weighted training and testing ('wRF' labels) for the six-class classification of the *Fermi*-LAT sources.

for all classes. The largest standard deviations of about 13 per cent are for the RF minus NN probabilities for the unweighted and weighted trainings in the bcu+ class.

Fig. 14 is similar to Fig. 13 but we divide the difference of the probabilities for each source by the RMS of the uncertainties due to random training/testing splits, i.e. we plot the histograms of $(p_1 - p_2)/\sqrt{(\sigma_1^2 + \sigma_2^2)/2}$. We see that in most cases the difference in class probabilities for the individual sources for different classification methods is comparable with the standard deviations due to training/testing splits (ranging from about 0.5σ to 2σ). We create probabilistic catalogues constructed both with unweighted and with weighted training samples.

In Fig. 15, we compare the confusion matrices for the RF classification with the weighted and unweighted training and testing data sets. The predicted classes are calculated by taking the class with the largest probability for each source. In this calculation, we take all associated sources and use the class probabilities determined as a mean over the training/testing splits when the sources are in

the testing data sets. We see that testing with unweighted samples gives similar performance estimates both for training on unweighted samples (top-left panel) and for training on weighted samples (top-right panel). While for testing with weighted samples, training with unweighted samples (bottom-left panel) has a better performance for psr+ and fsrq+ classes compared with the training with weighted samples (bottom-right panel).

6. CONCLUSIONS

In the paper, we study the effect of the covariate shift (due to difference in the distributions of associated and unassociated sources) on the multi-class classification of *Fermi*-LAT sources. In order to realistically estimate the expected performance for the classification of unassociated sources using only the associated sources, we introduce sampling weights proportional to the ratio of the PDFs for the unassociated sources to the associated sources, so that the
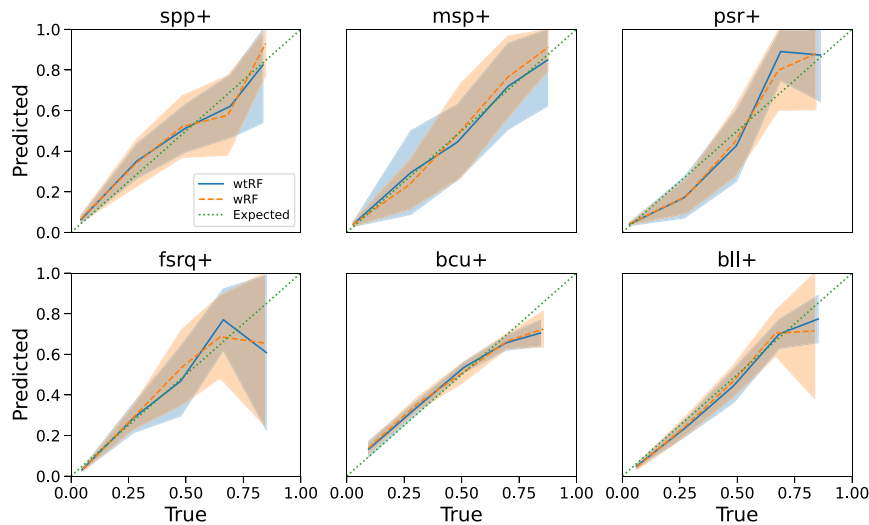
Figure 12. Reliability diagrams for unweighted training but weighted testing using the RF classification algorithm ('wtRF' labels) and weighted training and testing ('wRF' labels) for the six-class classification of the *Fermi*-LAT sources. The dotted line shows the optimal calibration of the predicted probabilities.

Table 3. Predictions for the number of associated and unassociated sources in the **weighted** training catalogue. For the definition of the classes, see Table 2.

	Class label	N assoc	RF assoc	NN assoc	RF unas	NN unas
1	spp+	173	183.0 ± 2.1	237.7 ± 2.8	406.7 ± 3.1	525.0 ± 3.5
2	msp+	179	183.4 ± 1.8	226.7 ± 4.0	177.1 ± 2.0	200.9 ± 2.4
3	psr+	241	238.8 ± 2.0	399.5 ± 5.3	187.5 ± 2.1	173.7 ± 2.1
4	fsrq+	827	792.5 ± 3.5	663.8 ± 5.7	197.5 ± 2.0	156.6 ± 1.9
5	bcu+	1637	1733.4 ± 4.7	1887.8 ± 6.3	1270.0 ± 4.1	1257.3 ± 4.1
6	bll+	1557	1482.8 ± 4.2	1198.5 ± 6.1	338.1 ± 2.5	263.4 ± 2.3

Table 4. Predictions for the number of associated and unassociated sources in the **unweighted** training catalogue. For the definition of the classes, see Table 2.

	Class label	N assoc	RF assoc	NN assoc	RF unas	NN unas
1	spp+	173	175.5 ± 2.0	170.7 ± 1.5	420.4 ± 3.1	448.9 ± 2.4
2	msp+	179	177.5 ± 1.7	181.3 ± 2.0	182.3 ± 2.0	197.0 ± 1.7
3	psr+	241	236.2 ± 2.0	238.8 ± 2.2	187.6 ± 2.1	181.7 ± 1.6
4	fsrq+	827	829.8 ± 3.5	828.3 ± 2.3	200.5 ± 2.1	185.7 ± 1.2
5	bcu+	1637	1641.8 ± 4.9	1619.0 ± 3.1	1253.3 ± 4.1	1269.7 ± 2.9
6	bll+	1557	1553.2 ± 4.2	1576.0 ± 3.0	332.9 ± 2.5	294.0 ± 1.6

PDF of associated sources weighted by these sampling weights is similar to the PDF of the unassociated ones.

We use RF and NN algorithms and perform training with both weighted and unweighted samples. We test the performance using weighted testing samples drawn from the associated sources, which is expected to give realistic estimates of the performance for the unassociated sources. We find that

(1) **Covariate shift has little effect on estimated class probabilities for individual sources.** The difference among class probabilities for individual unassociated sources derived with RF and NN algorithms using either weighted or unweighted training samples are comparable with the statistical uncertainties for the probabilities estimated from the random splits into training and testing data sets. This result justifies the use of unweighted training samples in the derivation of the classification algorithms, which are then used to classify unassociated sources.

(2) **Using weighted or unweighted training samples has little effect on average performance.** The average performance, estimated using ROC curves, precision, recall, and reliability diagrams, is similar for weighted and for unweighted training. However, the variance is observed to increase for NN algorithms in the weighted training case for some of the characteristics (e.g. precision and reliability) for classes with a significant decrease due to weights in the effective sample size (e.g. FSRQs).

(3) **Covariate shift results in a decrease of up to 20 per cent in precision and recall for some classes, estimated with weighted testing samples compared with estimates with unweighted testing samples.** The most affected classes are the classes of extragalactic sources (such as FSRQs and BL Lacs) dominated by sources at high latitudes where the fraction of unassociated sources is smaller than at low latitudes.

The overall conclusion is that both unweighted and weighted training have similar expected performance, but the covariate shift
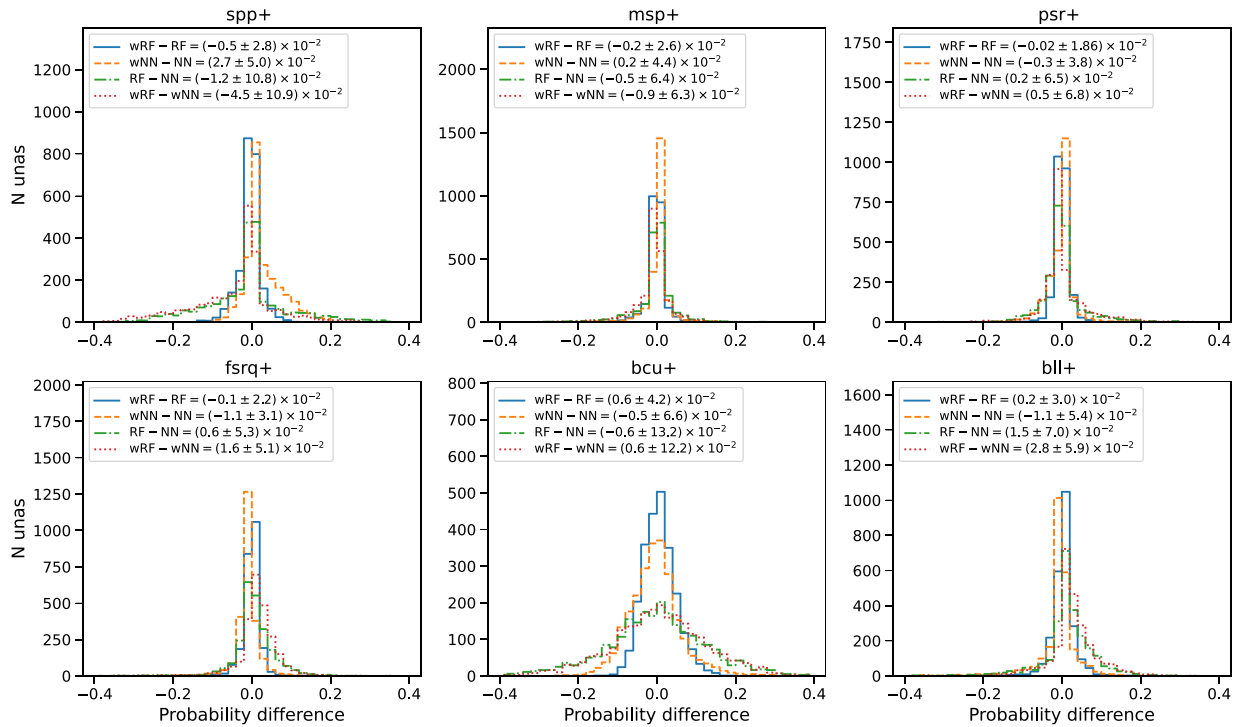
Figure 13. Difference of class probabilities for individual sources determined with RF ('RF' labels) and NN ('NN' labels) algorithms using weighted ('w' is included in labels) and unweighted training. The corresponding overall mean difference and the standard deviations are reported in the labels.
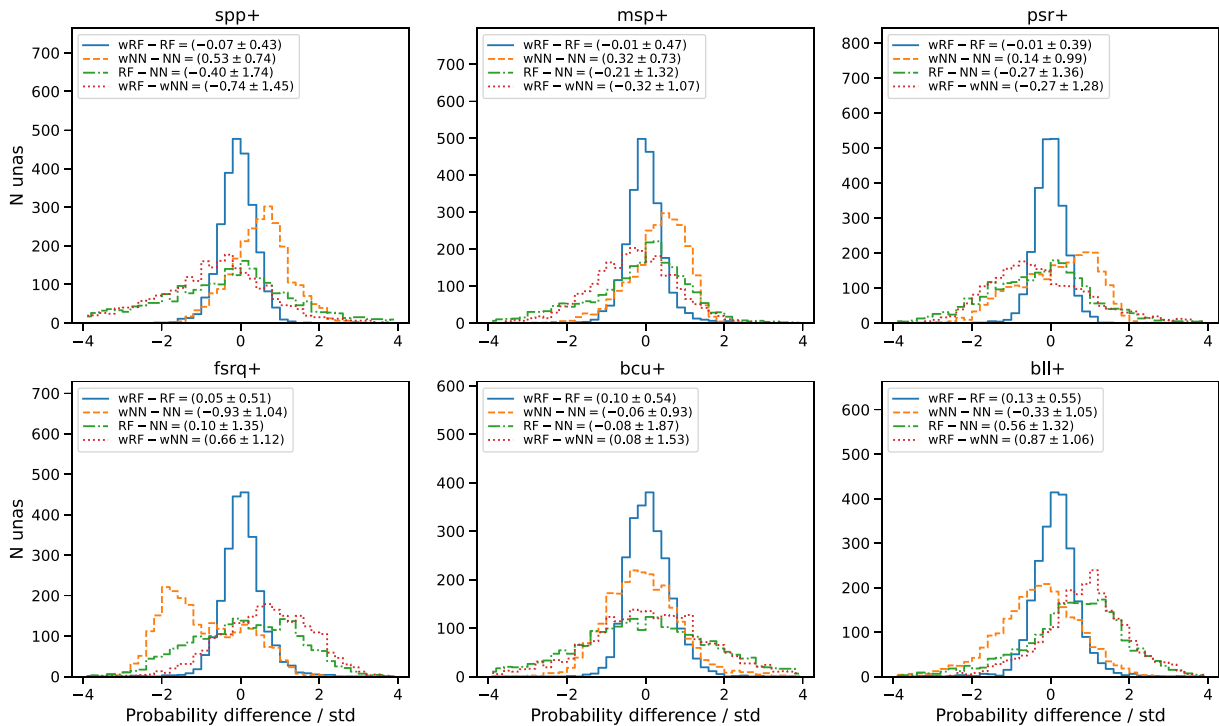

Figure 14. Difference of class probabilities for individual sources determined with RF ('RF' labels) and NN ('NN' labels) algorithms using weighted ('w' is included in labels) and unweighted training relative to the standard deviations of the predicted probabilities (see text for more details). The corresponding overall mean relative difference and the standard deviations of the relative differences are reported in the labels.

should be taken into account in the estimations of the performance, e.g. with the weighted testing samples. Weighted training can lead to larger variance of the expected performance (especially for classes with a significant reduction in effective sample number) and it also leads to biased predictions for the unweighted tests. As a result, we find that it is better to perform the classification with unweighted training data set but, for a realistic estimate of the classification performance, one should use weighted testing samples.
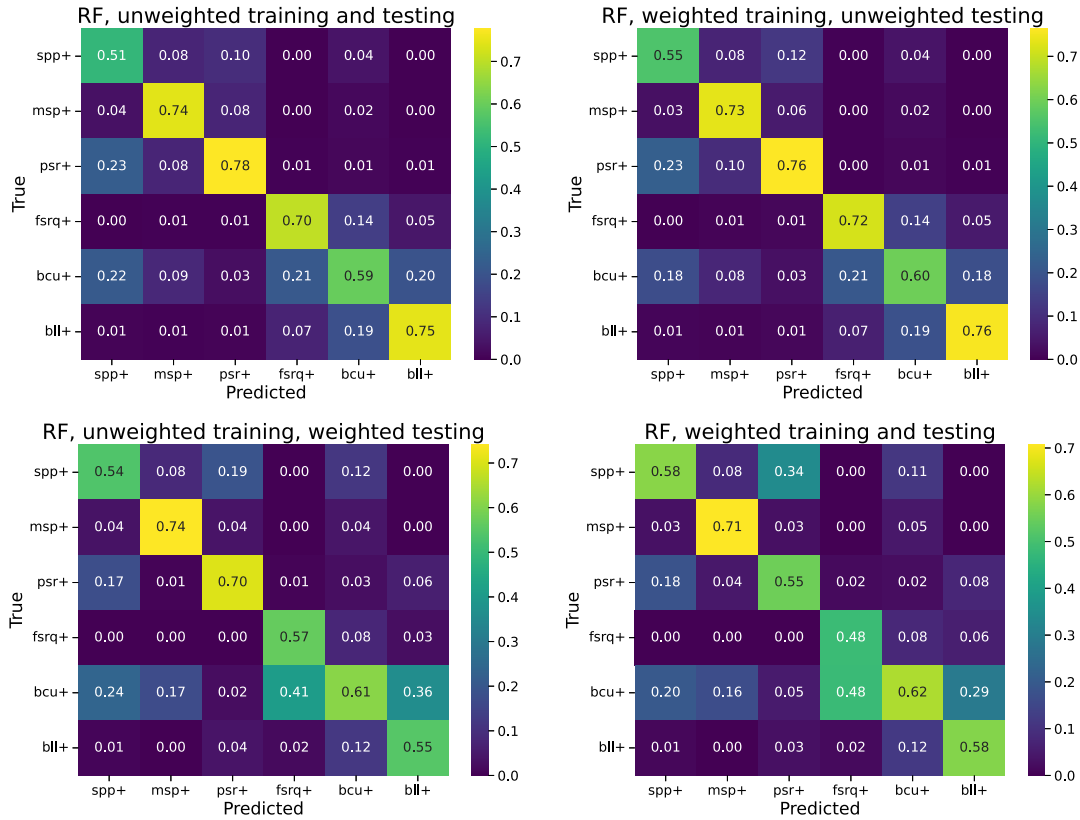
Figure 15. Confusion matrix normalized to the number of predicted sources in a class. The sources are classified according to the highest class probability for each source. The numbers on the diagonal show the one-versus-all precision for this classification method. Top (bottom) panels show the performance for the unweighted (weighted) samples representative for the associated (unassociated) sources.

We create probabilistic catalogues using RF and NN algorithms trained with weighted and unweighted samples. The catalogues include predicted class probabilities for six classes with RF and NN algorithms averaged over 51 realizations of the 70/30 per cent training/testing data sets (for associated sources the probabilities are averaged over the splits when the source appears in the testing sample) as well as the standard deviations of the probabilities due to the splits. We also add a column with the sample weights. We calculate the expected number of sources in the six classes among the unassociated sources. The largest fractional increase is expected for the spp+ class: There are about 2.5 times more expected spp+ sources among the unassociated ones than there are associated spp+ sources. For comparison, the expected number of msp+ (psr+) sources among the unassociated sources is about the number of (75 per cent of) associated msp+ (psr+) sources. For the extragalactic sources, the largest fractional increase is for the bcu+ sources: The expected increase is more than 70 per cent, compared with an increase of about 20 per cent for bll+ and fsrq+ classes. The catalogues are publicly available at https://zenodo.org/doi/10.5281/zenodo.8140548.

LEARN (https://scikit-learn.org/stable/, Pedregosa et al. 2011), and TENSORFLOW (https://www.tensorflow.org/, Abadi et al. 2015).

DATA AVAILABILITY

The results of this work are based on the publicly available *Fermi*-LAT 4FGL-DR4 catalogue https://fermi.gsfc.nasa.gov/ssc/data/access/lat/14yr_catalog/ (Ballet et al. 2023). The results of this work are available online at https://zenodo.org/doi/10.5281/zenodo.8140548.

REFERENCES

Abadi M. et al., 2015, TensorFlow: Large-scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org
Abdollahi S. et al., 2020, ApJS, 247, 33
Abdollahi S. et al., 2022, ApJS, 260, 53
Ackermann M. et al., 2012, ApJ, 753, 83
Akaike H., 1974, IEEE T. Autom. Cont., 19, 716
Ballet J., Bruel P., Burnett T. H., Lott B., The Fermi-LAT collaboration, 2023, preprint (arXiv:2307.12546)
Bhat A., Malyshev D., 2022, A&A, 660, A87
Finke T., Krämer M., Manconi S., 2021, MNRAS, 507, 4061
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Kish L., 1965, Survey Sampling. Wiley, New York
Lefaucheur J., Pita S., 2017, A&A, 602, A86
Luo S., Leung A. P., Hui C. Y., Li K. L., 2020, MNRAS, 492, 5377
McKinney W., 2010, Proc. 9th Python Sci. Conf., Data Structures for Statistical Computing in Python. p. 56
Malyshev D. V., Bhat A., 2023, MNRAS, 521, 6195

Mirabal N., Charles E., Ferrara E. C., Gonthier P. L., Harding A. K., Sánchez-Conde M. A., Thompson D. J., 2016, ApJ, 825, 69

Moreno-Torres J. G., Raeder T., Alaíz-Rodríguez R., Chawla N., Herrera F., 2012, Pattern Recognit., 45, 521

Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825

Robitaille T. P. et al., 2013, A&A, 558, A33

Saz Parkinson P. M., Xu H., Yu P. L. H., Salvetti D., Marelli M., Falcone A. D., 2016, ApJ, 820, 8

Schwarz G., 1978, Ann. Stat., 6, 461

Zhu K.-R., Kang S.-J., Zheng Y.-G., 2021, Res. Astron. Astrophys., 21, 015

SUPPORTING INFORMATION

Supplementary data are available at *RASTAI* online.

RASTI_v1_2023_10_11.zip

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

APPENDIX A: MODEL THE DISTRIBUTIONS OF ASSOCIATED AND UNASSOCIATED SOURCES

In this appendix, we construct PDFs of associated and unassociated sources using GMMs. In order to determine an optimal number of kernels in the GMMs, we compare the Akaike information criterion (AIC, Akaike 1974)

$$\mathrm{AIC} = 2k - 2\ln\mathcal{L}, \tag{A1}$$

where k is the number of parameters in a model and $\mathcal{L}$ is the likelihood of the model, and the Bayesian information criterion (BIC, Schwarz 1978)

$$\mathrm{BIC} = k\ln n - 2\ln\mathcal{L}, \tag{A2}$$

where k and $\mathcal{L}$ are the same as in the AIC and n is the number of data samples. In Fig. A1, we plot the AIC and BIC as functions of the number of the GMM kernels. We use the AIC and BIC implementations in the SCIKIT-LEARN package (Pedregosa et al. 2011). We see that above about 12 GMM kernels BIC is approximately constant for
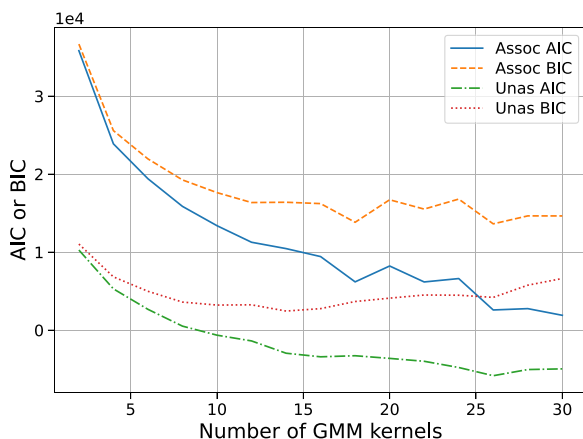
Figure A1. The Akaike (Akaike 1974) and Bayesian (Schwarz 1978) information criteria as functions of the number of GMM kernels for associated (unassociated) sources: 'Assoc AIC' and 'Assoc BIC' ('Unas AIC' and 'Unas BIC') labels, respectively. Both AIC and BIC are decreasing up to about 12 GMM kernels.
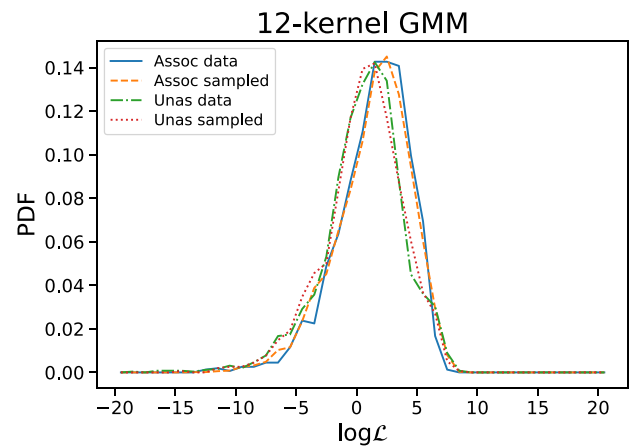
Figure A2. The distribution of log-likelihoods in the 12-kernel GMM models of the associated and unassociated sources. Solid (dash–dotted) line: The distribution of log-likelihoods for associated (unassociated) sources. Dashed (dotted) line: The distribution of log-likelihoods for sources sampled from the GMM model for associated (unassociated) sources.

both associated and unassociated sources. Although AIC continues to decrease above 12 kernels, we choose the GMM models with 12 kernels as the models with minimal complexity up to which both AIC and BIC are decreasing. We also test in Fig. A2 that for the 12-kernel GMM models the distributions of likelihoods for the associated and unassociated sources are similar to the distributions of likelihoods for the samples drawn from models for the associated and unassociated sources, respectively.

APPENDIX B: SELECTION OF INPUT FEATURES AND FEATURE IMPORTANCE

In this work, we use 10 input features (described in Section 2). Although *Fermi*-LAT catalogues have many more source parameters, most of these parameters are highly correlated (Bhat & Malyshev 2022). In particular, there are different representations of the same quantity, such as the position on the sky in Galactic or equatorial coordinates. There are also high correlations, e.g. among uncertainties in flux, energy flux, and spectrum normalization (Bhat & Malyshev 2022). It has been noted by Luo et al. (2020) that, in case of two-class classification, using more than five input features does not significantly improve the classification performance with the increasing complexity of the model.

In this work, we use the features previously selected by Luo et al. (2020) and Bhat & Malyshev (2022) with several modifications: (1) instead of the hardness ratios, we use the log-parabola curvature parameter to describe the change of the 'spectral index' as a function of energy; (2) instead of the spectral index parameter (or power-law spectral index), we use the index of the log-parabola spectrum at 1 GeV, which is independent of the pivot energy and is well defined for curved spectra; (3) we add the average significance parameter ('Signif_Avg'). Although this parameter is highly correlated with the energy flux above 100 MeV (Bhat & Malyshev 2022), 'log10(Signif_Avg)' has a higher importance than 'log10(Energy_Flux100)' in RF classifications with more than two classes (cf. Table B1); and (4) we replace Galactic longitude with two parameters 'cos(GLON)' and 'sin(GLON)' in order to avoid discontinuity between $0°$ and $360°$. The same input features have been previously used by Malyshev & Bhat (2023) in the determination of the hierarchical classification of the *Fermi*-LAT sources.

Table B1. Feature importance for RF classification with unweighted training and testing for different numbers of classes corresponding to the different depth of the class separation in the bottom panel of Fig. 2. Depths 1, 2, 3, and 4 correspond to two-, four-, five-, and six-class classifications, respectively.

Feature	Depth 1	Depth 2	Depth 3	Depth 4
LP_index1000MeV	0.159	0.167	0.160	0.163
log10(Signif_Avg)	0.077	0.128	0.140	0.142
log10(Variability_Index)	0.100	0.114	0.114	0.113
log10(Unc_Energy_Flux100)	0.128	0.109	0.105	0.108
LP_beta	0.103	0.106	0.105	0.101
log10(Energy_Flux100)	0.112	0.089	0.092	0.092
sin(GLAT)	0.055	0.081	0.087	0.087
LP_SigCurv	0.164	0.097	0.082	0.081
cos(GLON)	0.052	0.056	0.058	0.059
sin(GLON)	0.051	0.054	0.057	0.055

We show the importance of the 10 input features in the RF classification (in the unweighted training case) in Table B1. The features are ordered according to the importance for the six-class classification (the 'Depth 4' column). It is interesting to note that some features have a significantly different importance in the two-class and in the multi-class classifications. For instance, the spectral curvature significance parameter ('LP_SigCurv') has the highest significance for the two-class classification (it has also been the most significant parameter in the analysis of Saz Parkinson et al. 2016, Luo et al. 2020, and Bhat & Malyshev 2022), but it is in the middle (near the end) of the list for the four-class (five- and six-class) classification. The parameter 'LP_index1000MeV' is the most important parameter for all classifications in this work, while 'Spectral_Index' has been less significant in the previous two-class classifications, e.g. on place three in Luo et al. (2020) or on place four in Saz Parkinson et al. (2016), and log-parabola index 'LP_Index' was near the end of the significance table in Malyshev & Bhat (2023). On the other hand, parameters 'log10(Variability_Index)' and 'log10(Unc_Energy_Flux100)' have similar importance for all cases (places three and four, respectively) in this work as well as in the previous analyses, where the importance of these features has been between place two and four (Saz Parkinson et al. 2016; Luo et al. 2020; Bhat & Malyshev 2022).

APPENDIX C: NEURAL NETWORKS

In this appendix, we provide details about the NN method for the classification of *Fermi*-LAT sources. We use the TensorFlow implementation of NNs (Abadi et al. 2015). We use stochastic gradient descent (adam) with learning rate of 0.001, two hidden layers with 20 and 10 nodes, respectively, tanh activation functions, batch size of 200, L2 regularization with $l2 = 0.001$, and no drop out. We use the sparse categorical cross entropy loss function. In Fig. C1, we show the one-versus-all ROC AUC values for the six groups for the unweighted (top panel) and weighted (bottom panel) training. We find that for the unweighted training there is no sign of overfitting up to the maximal number of epochs used in this test, e.g. 2000. Provided that there is still a slight increase in performance in some groups up to about 500 epochs, we use 500 epochs for training the NN algorithm in the unweighted case. For the weighted case, there are signs of overfitting for some groups above 500 epochs, e.g. for psr+ and fsrq+ groups. As a result, we also use 500 epochs for the training in the weighted samples case.

In Fig. C2, we compare ROC curves, precision, recall, and calibration diagrams for the training with weighted samples using

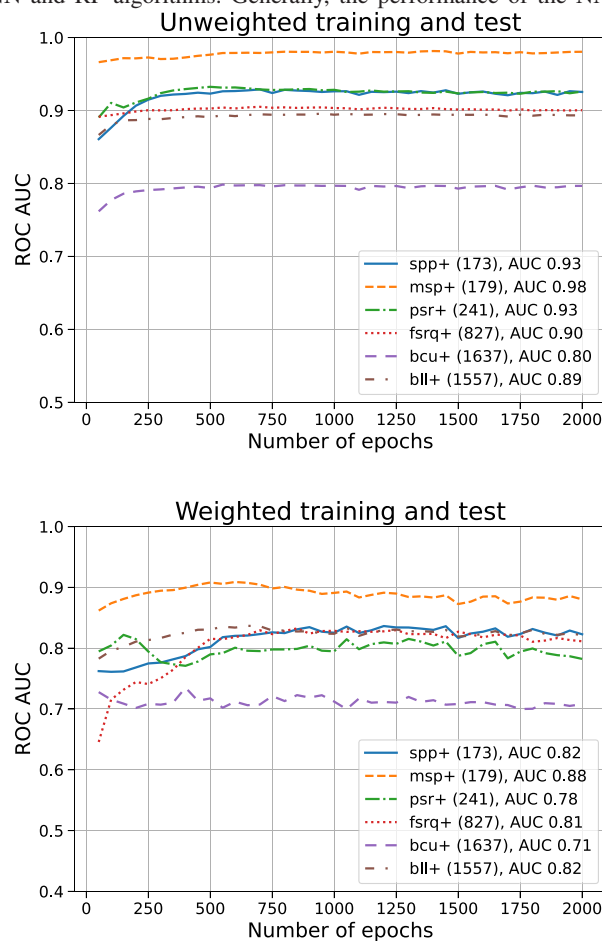NN and RF algorithms. Generally, the performance of the NN is

Figure C1. The ROC AUC for one-versus-all classification as a function of the number of epochs for the NN algorithm with unweighted (top panel) and weighted (bottom panel) samples.

comparable with the performance of the RF. RF gives slightly better results for the psr+ class, while NN has a better performance for the bll+ class.

In Fig. C3, we compare the performance of the NN algorithm trained with unweighted and tested with weighted samples ('wtNN' labels) versus trained and tested with weighted samples ('wNN' labels). Most of the characteristics are similar for training with weighted and unweighted samples. However, the statistical uncertainty band is narrower in the unweighted training case for the fsrq+ and bll+ classes for precision (middle panels) and reliability (bottom panels). This can be attributed to the fact that in the weighted samples case the ratio of the effective number of samples to the number of associated sources in the fsrq+ and bll+ classes is the smallest among the six classes, which leads to an increased variance for this class in the weighted training case compared with the unweighted training.

In Fig. C4, we compare the confusion matrices for the weighted and unweighted training and testing data sets for the classification with the NN algorithm similar to the confusion matrices in Fig. 15 for the classification with the RF algorithm. The training on unweighted samples (left panels) has a similar or better performance (with a few exceptions) than the training on weighted samples (right panels) when tested both on unweighted samples (top panels) and on weighted samples (bottom panels).
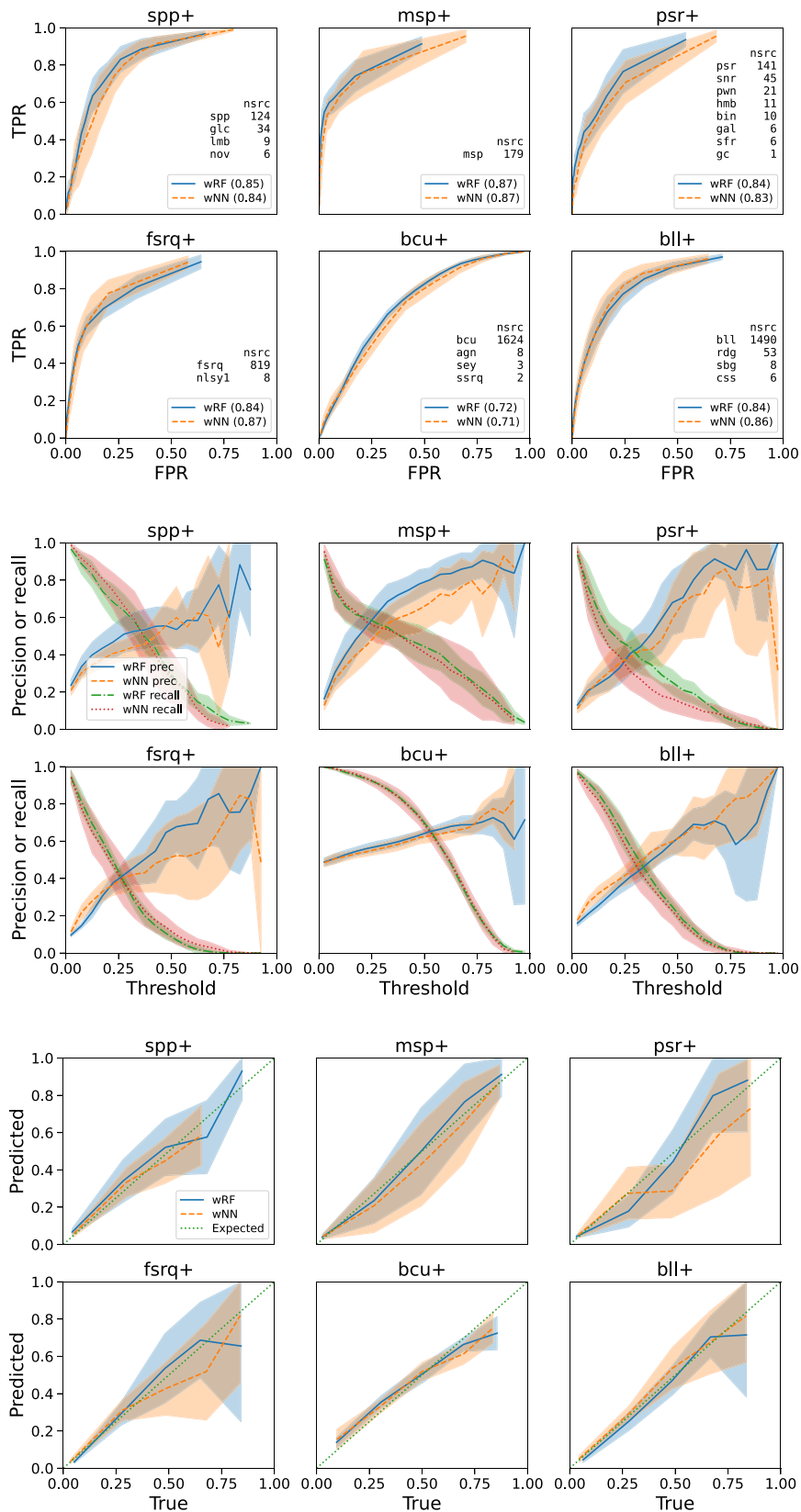
Figure C2. Comparison of ROC curves (top panels), precision and recall (middle panels), and calibration diagrams (bottom panels) for training and testing with weighted samples using RF ('wRF' labels) and NN ('wNN' labels) algorithms. The numbers in parenthesis in the top panel show the ROC AUC.
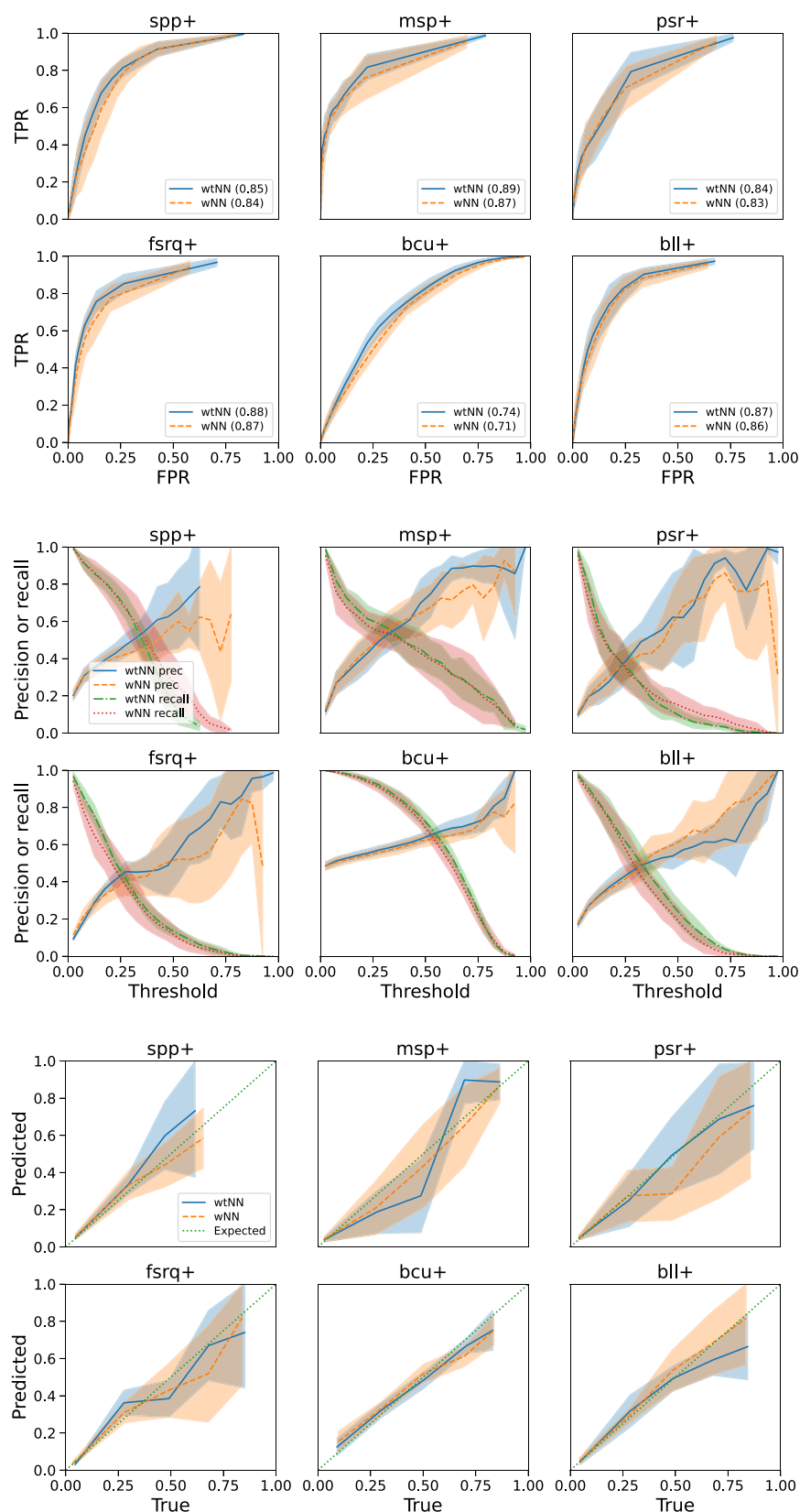
Figure C3. Comparison of ROC curves (top panels), precision and recall (middle panels), and calibration diagrams (bottom panels) for classification with NN algorithms in case of training with unweighted samples and testing with weighted samples ('wtNN' labels) and with weighted samples used both for training and testing ('wNN' labels).
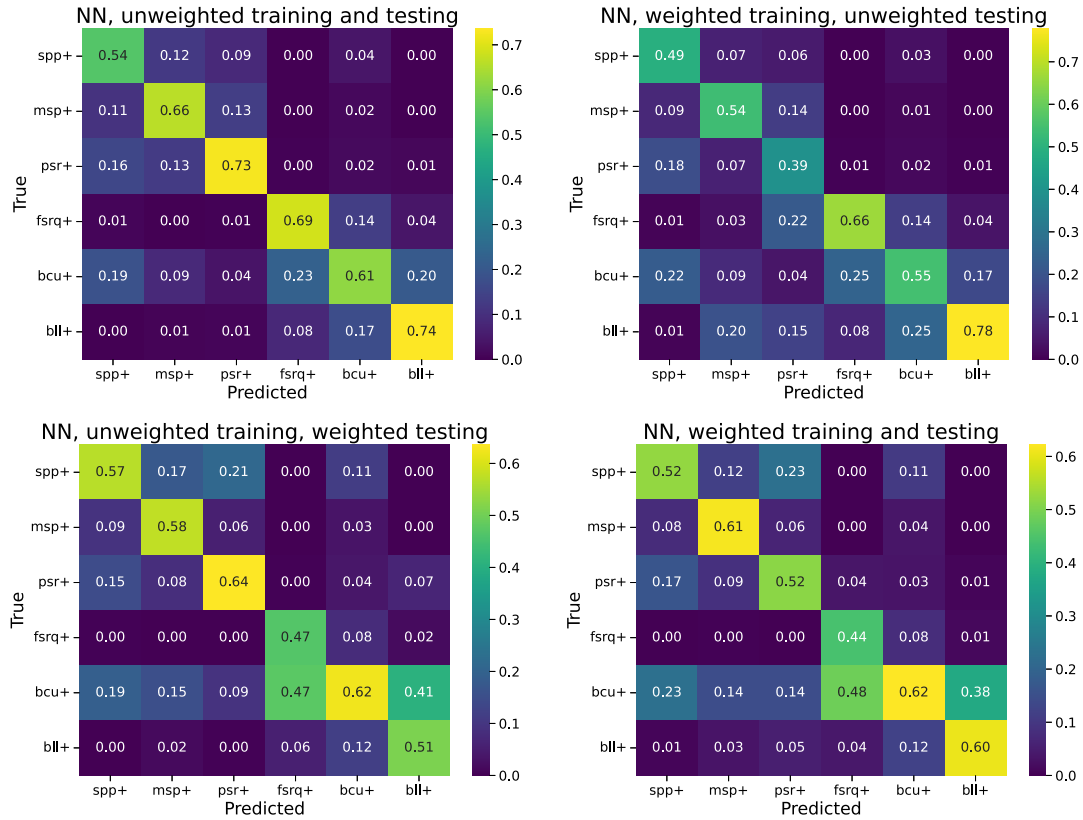
Figure C4. Confusion matrix normalized to the number of predicted sources in a class for the NN classification. The sources are classified according to the highest class probability for each source. The numbers on the diagonal show the one-versus-all precision for this classification method. Top (bottom) panels show the performance for the unweighted (weighted) samples representative for the associated (unassociated) sources.

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