PROCEEDINGS OF SCIENCE

# Towards Refined Population Studies: High-Confidence Blazar Candidates and their Multiwavelength Counterparts using Machine Learning

**Sabrina Einecke***
*TU Dortmund*
*E-mail:* sabrina.einecke@tu-dortmund.de

**Dominik Elsässer**
*TU Dortmund*
*E-mail:* dominik.elsaesser@tu-dortmund.de

**Wolfgang Rhode**
*TU Dortmund*
*E-mail:* wolfgang.rhode@tu-dortmund.de

**Katharina Morik**
*TU Dortmund*
*E-mail:* katharina.morik@tu-dortmund.de

The Third *Fermi*-LAT source catalog (3FGL) presents a large number of gamma-ray point sources affiliated with source types and counterparts. Nonetheless, 1010 sources remain unassociated and 573 sources are associated with active galaxies of uncertain type. Assigning blazar classes to these unassociated and uncertain sources, and linking counterparts to the unassociated ones, will refine tremendously our knowledge of the population of gamma-ray emitting objects.
To figure out the most likely counterpart, the sample of associated 3FGL sources is used to train machine learning classification algorithms. For any particular 3FGL source, all possible combinations with measurements of one additional energy range are considered, e. g. from the Wide-Field Infrared Survey Explorer (WISE) source catalog, the catalog of Faint Images of the Radio Sky at Twenty cm (FIRST), or the *Swift* X-ray Point Source (1SXPS) catalog. By merging the most probable candidates of each of those studies, the power of multiwavelength strategies is exploited and conclusions with even higher confidence concerning blazar counterpart candidates are drawn. In this contribution, the statistical model and its validation to estimate the performance is described. Finally, results of the application of this novel wavelength-dependent approach are presented, and its consequences concerning blazar population studies are discussed.

*35th International Cosmic Ray Conference*
*10-20 July, 2017*
*Bexco, Busan, Korea*

*Speaker.





PoS(ICRC2017)626

## 1. Introduction

The deepest all-sky survey in gamma rays has been performed with the Large Area Telescope (LAT) on board the *Fermi* satellite. The corresponding catalog of four years of observation – the 3FGL catalog [1] – comprises 3033 point sources. Out of these, 1010 sources remain without plausible associations, and 573 sources are associated to active galaxies of uncertain type. The assignment of blazar classes to these unassociated and uncertain sources, and the link of counterparts to the unassociated ones, is important to refine our knowledge of the population of gamma-ray emitting objects.

In recent years, the application of machine learning algorithms has become an important part in the exploration of astrophysical data. Previous machine learning approaches for the assignment of source types were based solely on properties extracted from gamma-ray observations. A binary classification between Active Galactic Nuclei (AGN) and pulsars has been performed by Ackermann et al. [2] and Mirabal et al. [3]. While Ackermann et al. applied a logistic regression and classification trees to the 1FGL catalog, Mirabal et al. used random forests and the 2FGL catalog. This catalog has also been investigated with machine learning strategies by Doert and Errando [4] and Hassan et al. [5]. Doert and Errando discriminated AGN and non-AGN with a combination of random forests and neural networks, and, moreover, Hassan et al. studied the application of random forests and support vector machines for the classification of the AGN subclasses BL Lacs (BLLs) and Flat Spectrum Radio Quasars (FSRQs).

For a better classification, additional source type-specific features are acquired by the extension to multiwavelength information. The prospects of adding multiwavelength information to the gamma-ray properties of a source have been extensively investigated by Massaro et al. (e. g., [6] and [7]), and its importance has been proven. This extension also provides the possibility to determine the most likely corresponding counterpart.

## 2. Data Samples

The 3FGL catalog comprises 3033 gamma-ray point sources, measured with the *Fermi*-LAT in an energy regime between 100 MeV and 300 GeV. The most numerous associations amongst the sources are BLLs (660) and FSRQs (484). Correspondingly, a classification of blazar types is particularly promising. Results from eight years of observations with the X-Ray Telescope (XRT) of the *Swift* satellite are included in the *Swift*-XRT Point Source Catalog (1SXPS) [8]. A number of 151 524 point sources have been detected in the energy regime of the telescope of 0.3-10 keV. The positional accuracy is approximately 5.5 arcseconds. The Wide-field Infrared Survey Explorer (WISE) conducted a mid-infrared survey at the wavelengths 3.4, 4.6, 12 and 22 $\mu$m (56-365 meV) [9]. The appertaining ALLWISE Source Catalog comprises 747 634 026 sources. The positional accuracy is approximately 6 arcseconds. The catalog of Faint Images of the Radio Sky at Twenty-centimeters (FIRST) [10] incorporates 946 432 radio sources, compiled over 18 years of observations with the Very Large Array at a wavelength of 20 cm (6.2 $\mu$eV). In contrast to the above-mentioned surveys, this survey only covers the north and south Galactic caps. The positional accuracy is approximately 1 arcsecond at 90% confidence. All catalogs provide information, amongst others, about fluxes, source positions, variabilities, significances, and spectral properties.

Since the positional accuracy of 3FGL point sources is in the order of several arcminutes, several hundred possible counterparts might be located within the uncertainty region of one particular 3FGL source, making the counterpart association ambiguous. This demonstrates the challenges of such an association procedure and the need for an advanced procedure. The contemplated approach is the consideration of all possible combinations of a specific 3FGL source and every counterpart candidate of one of the catalogs, located within the confidence region of the 3FGL source.

For the creation of the training sample (to create the classification model, see Section 3), the 3FGL sources with known positions of the associated counterpart are taken into account. The possible counterparts from the ALLWISE, 1SXPS and FIRST catalogs, respectively, are added. While the closest counterpart (with an adapted maximum distance) is associated with the classes BLL or FSRQ (depending on the association within the 3FGL catalog), the remaining counterparts are assigned the class non-Blazar.

By analogy, two further samples are constructed: One comprising the unassociated 3FGL sources, and the other including the AGNs of uncertain type. To these samples, the classification models will be applied.

## 3. Methods

The classification procedures have been performed with the machine learning library `scikit-learn` (*http://scikit-learn.org/*) and further libraries for the programming language python. Subsequent to the preparation of the data samples, typical machine learning approaches start with a feature generation and a feature selection. Features have been generated combining the different catalogs, e. g. colors, hardness ratios, and hardness slopes, equivalent to approaches used by Massaro et al. [6] and Ackermann et al. [2]. A feature selection has been conducted to remove unnecessary features and to minimize systematic uncertainties. First, features with a high absolute Spearman's rank correlation coefficient have been removed. Then, a recursive feature elimination has been performed with the same classifier and settings that are used for the final classification models. This method trains a classification model based on all features, assigns ranks to them, and removes the ones with the lowest ranks. Based on this new feature set, a new classification model is trained, and the procedure is repeated recursively. For all models, 30 features have been selected. A robust performance dependent on the number of selected features has been proven by evaluating the performance score Area Under the Curve (AUC, see below for an explanation).

The presented problem is a multiclass problem, since the classification task is to discriminate the three classes BLL, FSRQ and non-Blazar. As a classifier a random forest has been chosen with 200 trees, the information gain as split selection criteria, 5 features considered at each node, and a minimum sample size of 5 for a split. The random forest model assigns specific class scores (BLL or FSRQ score) to every source (or counterpart combination) – the higher the score, the more likely the source is associated to this class.

The performance of the model is estimated with a 10-fold cross validation. The training sample is split with a ratio of 9:1 – one set is used to train the model, the other one to assign a class to each source and to compare it with the true class. The performance is quantified by different values, such as purity (the fraction of correctly classified events of a specific class to all events of this class), efficiency (the ratio of correctly classified events and the total number of events of the class), or

AUC (the area under the curve, spanning the efficiency of one class against the others). Through an iteration of the split sets, these performance values are calculated multiple times, and a mean and standard deviation is derived.

## 4. Performance

For each catalog (ALLWISE, 1SXPS and FIRST) of a different wavelength, classification models have been trained, and their performances have been evaluated. The distributions of the BLL and FSRQ scores are shown in Figures 1-3. The ALLWISE model (mid-infrared) reaches AUCs of $0.995 \pm 0.007$ for BLLs and $0.99 \pm 0.01$ for FSRQs. While the 1SXPS model (X-ray) obtains AUCs of $0.97 \pm 0.01$ for BLLs and $0.96 \pm 0.02$ for FSRQs, the FIRST model (radio) achieves AUCs of $0.977 \pm 0.008$ for BLLs and $0.987 \pm 0.006$ for FSRQs.

To get an impression of the purities and efficiencies of the models, exemplary score thresholds have been chosen such that purities of approximately 0.8 are achieved. For the ALLWISE model, a purity of $0.81 \pm 0.08$ and an efficiency of $0.85 \pm 0.04$ is obtained for an exemplary BLL score threshold of 0.3, and a purity of $0.82 \pm 0.08$ and an efficiency of $0.52 \pm 0.08$ is obtained for an exemplary FSRQ score threshold of 0.5. For the 1SXPS model, a purity of $0.78 \pm 0.06$ and an efficiency of $0.80 \pm 0.06$ is obtained for an exemplary BLL score threshold of 0.4, and a purity of $0.84 \pm 0.08$ and an efficiency of $0.58 \pm 0.08$ is obtained for an exemplary FSRQ score threshold of 0.6. For the FIRST model, a purity of $0.8 \pm 0.2$ and an efficiency of $0.16 \pm 0.08$ is obtained for an exemplary BLL score threshold of 0.8, and a purity of $0.8 \pm 0.1$ and an efficiency of $0.6 \pm 0.1$ is obtained for an exemplary FSRQ score threshold of 0.6.

By merging the resulting candidates from the models described above, the performance is improved. The combination of the results obtained with the ALLWISE and FIRST catalog leads to a purity of 0.86 for BLLs and 0.89 for FSRQs, when constraining the distance between the counterparts to 7 arcseconds. A purity of 0.90 for both BLLs and FSRQs is achieved for the combination of the ALLWISE and the 1SXPS catalog, and a largest distance between the counterparts of 11.5 arcseconds.

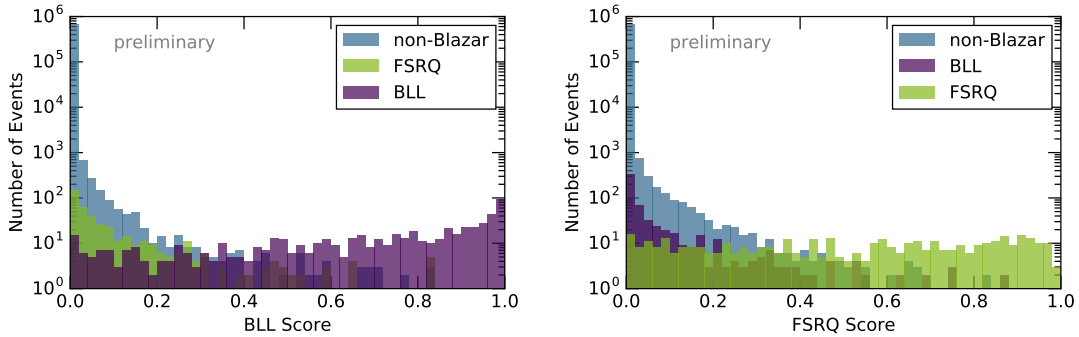


**Figure 1:** Distribution of the BLL (*left*) and the FSRQ (*right*) scores, derived with a Random Forest classification model. The model has been applied to the training sample, created with the 3FGL and the ALLWISE catalogs.

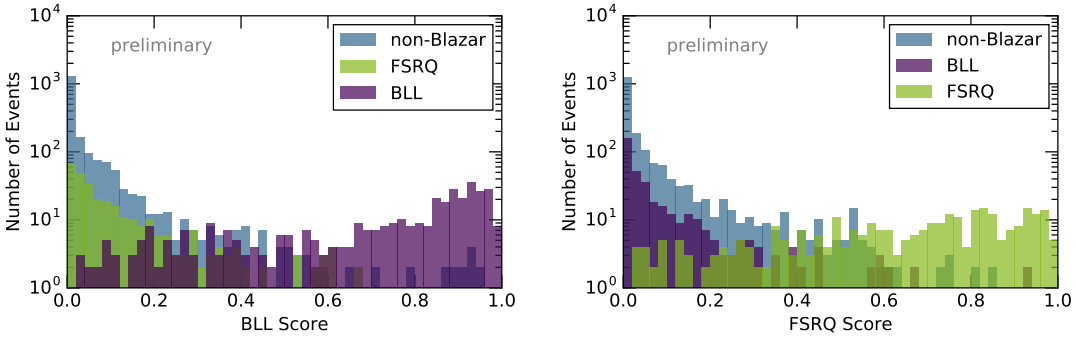


**Figure 2:** Distribution of the BLL (*left*) and the FSRQ (*right*) scores, derived with a Random Forest classification model. The model has been applied to the training sample, created with the 3FGL and the 1SXPS catalogs.

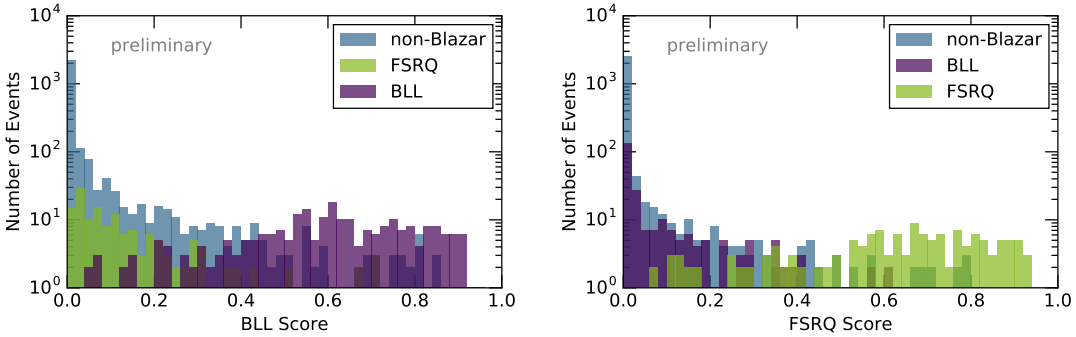


**Figure 3:** Distribution of the BLL (*left*) and the FSRQ (*right*) scores, derived with a Random Forest classification model. The model has been applied to the training sample, created with the 3FGL and the FIRST catalogs.

## 5. Results

The application of the classification models to the samples of unassociated sources and AGNs of uncertain type leads to distributions of the BLL and FSRQ scores as shown in Figures 4-6. For the ALLWISE model, 340 BLL and 30 FSRQ candidates have been found for the unassociated sample and exemplary score thresholds of 0.3 and 0.5, respectively, and 232 BLL and 45 FSRQ candidates for the sample of AGNs of uncertain type. For the 1SXPS model, 35 BLL and 1 FSRQ candidates have been found for the unassociated sample, and 90 BLL and 18 FSRQ candidates for the sample of AGNs of uncertain type and exemplary score thresholds of 0.4 and 0.6, respectively. For the FIRST model, 18 BLL and 0 FSRQ candidates have been found for the unassociated sample and exemplary score thresholds of 0.6 and 0.8, respectively, and 12 BLL and 1 FSRQ candidates for the sample of AGNs of uncertain type.

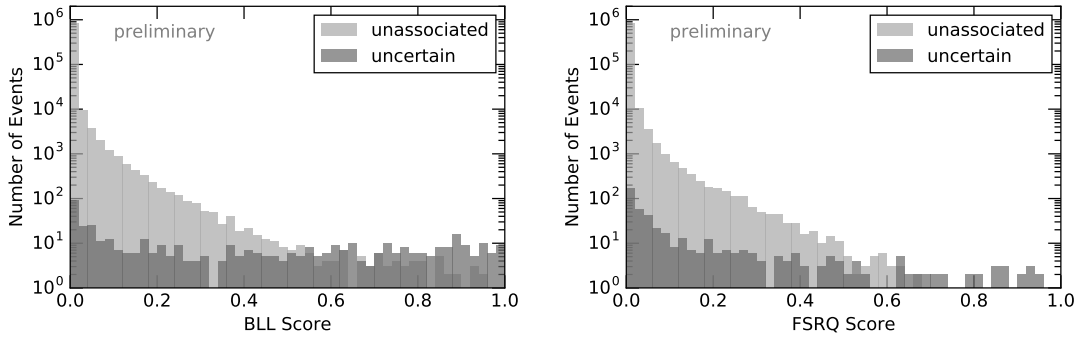


**Figure 4:** Distribution of the BLL (*left*) and the FSRQ (*right*) scores, derived with the above-mentioned Random Forest classification model. The model has been applied to the unassociated sample and the sample of AGNs of uncertain type, created with the 3FGL and the ALLWISE catalogs.

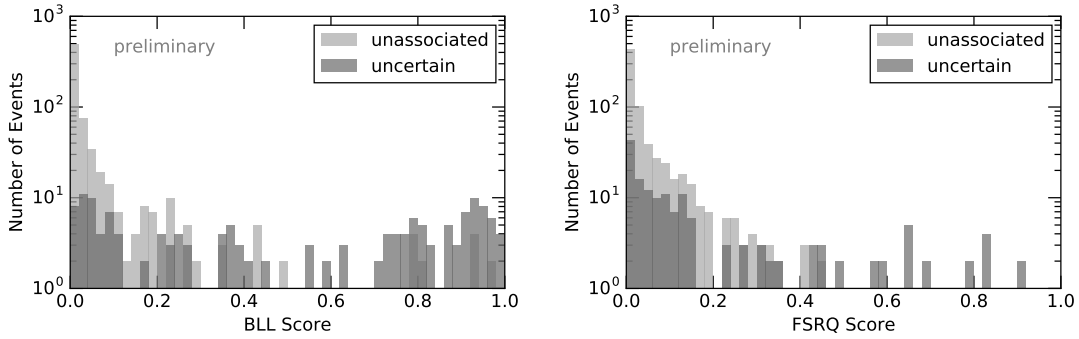


**Figure 5:** Distribution of the BLL (*left*) and the FSRQ (*right*) scores, derived with the above-mentioned Random Forest classification model. The model has been applied to the unassociated sample and the sample of AGNs of uncertain type, created with the 3FGL and the 1SXPS catalogs.

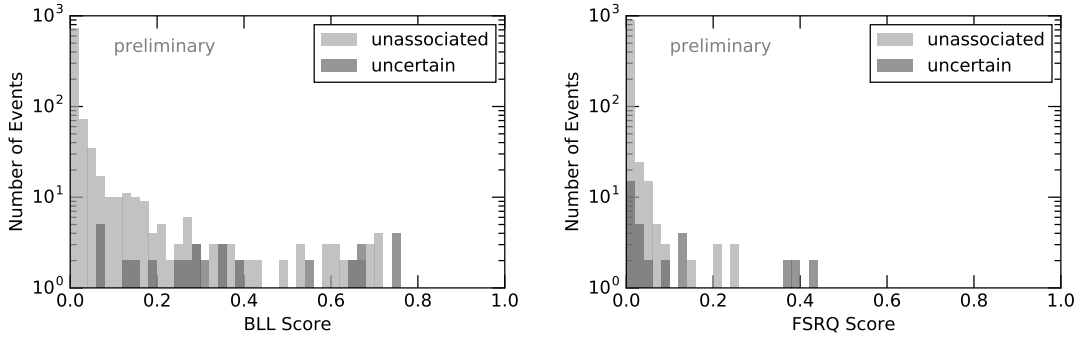


**Figure 6:** Distribution of the BLL (*left*) and the FSRQ (*right*) scores, derived with the above-mentioned Random Forest classification model. The model has been applied to the unassociated sample and the sample of AGNs of uncertain type, created with the 3FGL and the FIRST catalogs.

Merging the candidates resulting from the application of the models to the unassociated sample, 20 BLL and 0 FSRQ candidates have been found for the ALLWISE/1SXPS combination, and 8 BLL and 0 FSRQ candidates for the ALLWISE/FIRST combination. The application of the models to the sample of AGNs of uncertain type leads to 69 BLL and 13 FSRQ candidates for the ALLWISE/1SXPS combination, and 9 BLL and 0 FSRQ candidates for the ALLWISE/FIRST combination with high confidence.

## 6. Conclusion

Based on the 3FGL catalog of gamma-ray point sources and additional catalogs (ALLWISE, 1SXPS and FIRST) of different wavelengths (mid-infrared, X-ray and radio), random forest classification models have been trained to assign blazar classes, and to link counterparts at the same time. The performance of these models have been evaluated with a cross-validation, and AUC performance values between 0.96 and 0.99 could have been achieved for the individual models. Applying these models to the unassociated 3FGL sources and the AGNs of uncertain type, led to the obtainment of blazar candidates with their corresponding multiwavelength counterparts. The confidence of these candidates is increased by the combination of the results of the individual models, and high-confidence blazar candidates are obtained. The promising results prove the prospects and capabilities of this newly developed method. The findings could be deployed for population studies and dark matter searches, but could also help to resolve the origin of astrophysical neutrinos or ultra-high-energy cosmic rays.

## Acknowledgements

Part of this work is supported by Deutsche Forschungsgemeinschaft (DFG). This work made use of data supplied by the UK Swift Science Data Centre at the University of Leicester. This publication makes use of data products from the Wide-field Infrared Survey Explorer, which is a joint project of the University of California, Los Angeles, and the Jet Propulsion Laboratory/California Institute of Technology, and NEOWISE, which is a project of the Jet Propulsion Laboratory/California Institute of Technology. WISE and NEOWISE are funded by the National Aeronautics and Space Administration. This research has made use of the NASA/IPAC Infrared Science Archive, which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

## References


[1] F. Acero et al. *Fermi Large Area Telescope Third Source Catalog. ApJS* **2015**, *218*, 2.

[2] M. Ackermann et al. A Statistical Approach to Recognizing Source Classes for Unassociated Sources in the First *Fermi*-LAT Catalog. *ApJ* **2012**, *753*, 83.

[3] N. Mirabal et al. *Fermi's SIBYL: Mining the Gamma-Ray Sky for Dark Matter Subhalos. MNRAS* **2012**, *424*, L64.

[4] M. Doert and M. Errando *Search for Gamma-Ray-Emitting Acitve Galactic Nuclei in the Fermi-LAT unassociated Sample using Machine Learning. ApJ* **2014**, *782*, 41.

[5] T. Hassan et al. *Gamma-Ray Active Galactic Nucleus Type through Machine-learning algorithms. MNRAS* **2013**, *428*.

[6] F. Massaro et al. *The WISE Gamma-Ray Strip Parametrization: The Nature of the Gamma-Ray Active Nuclei of Uncertain Type. ApJ* **2012**, *750*, 138.

[7] A. Paggi et al. *Unveiling the Nature of the Unidentified Gamma-Ray Sources. IV. The Swift Catalog of Potential X-Ray Counterparts. ApJS* **2013**, *209*, 9.

[8] P. A. Evans et al. *1SXPS: A deep Swift X-Ray Telescope Point Source Catalog with Light Curves and Spectra. ApJS* **2014**, *210*, 1.

[9] E. L. Wright et al. *The Wide-Field Infrared Survey Explorer (WISE): Mission Description and Initial On-Orbit Performance. AJ* **2010**, *140*, 6.

[10] D. J. Helfand, R. L. White and R. H. Becker *The Last of FIRST: The Final Catalog and Source Identifications ApJ* **2015**, *801*, 1.