Finding active galactic nuclei through Fink

Etienne Russeil
Université Clermont Auvergne
CNRS/IN2P3, LPC
F-63000 Clermont-Ferrand, France
etienne.russeil@uca.fr

Emille E. O. Ishida
Université Clermont Auvergne
CNRS/IN2P3, LPC
F-63000 Clermont-Ferrand, France
emille.ishida@clermont.in2p3.fr

Roman Le Montagner
Université Paris-Saclay
CNRS/IN2P3, IJCLab
91405 Orsay, France
roman.le-montagner@ijclab.in2p3.fr

Julien Peloton
Université Paris-Saclay
CNRS/IN2P3, IJCLab
91405 Orsay, France
peloton@ijclab.in2p3.fr

Anais Möller
Centre for Astrophysics and Supercomputing
Swinburne University of Technology
Mail Number H29, PO Box 218, 31122, Hawthorn, VIC, Australia
amoller@swin.edu.au

Abstract

We present the Active Galactic Nuclei (AGN) classifier as currently implemented within the `Fink` broker. Features were built upon summary statistics of available photometric points, as well as color estimation enabled by symbolic regression. The learning stage includes an active learning loop, used to build an optimized training sample from labels reported in astronomical catalogs. Using this method to classify real alerts from the Zwicky Transient Facility (ZTF), we achieved 98.0% accuracy, 93.8% precision and 88.5% recall. We also describe the modifications necessary to enable processing data from the upcoming Vera C. Rubin Observatory Large Survey of Space and Time (LSST), and apply them to the training sample of the Extended LSST Astronomical Time-series Classification Challenge (ELAsTiCC). Results show that our designed feature space enables high performances of traditional machine learning algorithms in this binary classification task.

1 Introduction

Active Galactic Nuclei (AGN) are bright, variable astrophysical sources associated with the inflow of circumstellar matter into central galactic black holes [1]. From the observer perspective, they comprise a large set of light curve behaviors, including instances where observational patterns evolve with time [2, 3]. Beyond being paramount for the study of accretion and photoionization physics [4], they can trace star formation regions [5] and have the potential to enrich cosmological studies [6]. Thus, reliable and cheap identification of large populations of AGNs are crucial to enable a better understanding of their mechanisms and their impact in the galactic environment.

The Vera C. Rubin Observatory Legacy Survey of Space and Time[1] (LSST), expected to start operations in 2024, will produce a large volume of photometric data, including a diverse AGN

[1] `https://www.lsst.org/`

Machine Learning and the Physical Sciences workshop, NeurIPS 2022.

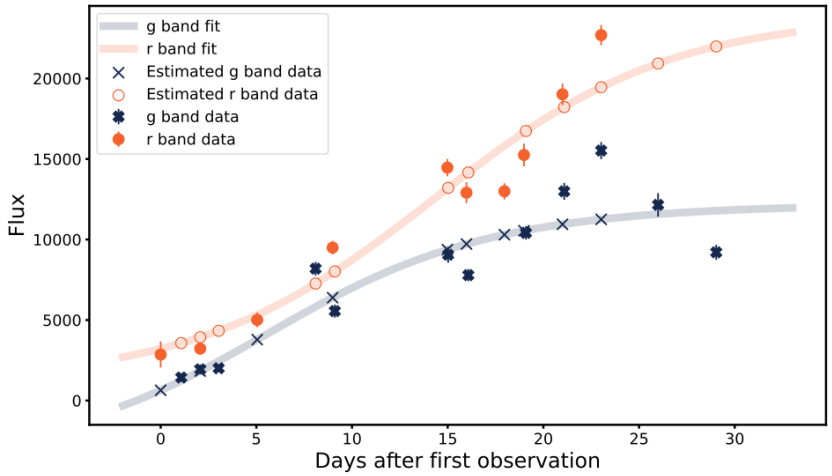

Figure 1: Example of `Fink-data` light curve. The object belongs to the BLLac class, with a ZTF objectID = "ZTF18abzwaiw" and an alert ID = 1424109966315015005. Bold symbols show observed points. The continuous lines show flux estimation using Equation 1 and open symbols denote flux value used for color estimation. ZTF filters g and r are shown in dark blue and orange, respectively.

population. Each time a brightness variability beyond 5-σ from the background is detected, an alert will be generated. We expect 10 million of such alerts per night, which will be streamed to chosen community brokers. `Fink`[2], is one of the official LSST brokers, whose task is to receive this data, extract meaningful information from it and re-distribute it to scientific communities. The broker contains a series of machine learning based modules, which enables fast processing of the data stream. In preparation for the start of LSST, `Fink` is currently operating on data from the Zwicky Transient Facility (ZTF) ([7]), which produces around 300 000 alerts per night.

This work presents details about the AGN classifier within the `Fink` broker. It includes a tailored feature extraction procedure followed by the construction of an optimized training sample using uncertainty sampling active learning and a random forest classifier [8].

2 Data

Two databases were used in this work. `Fink-data` corresponds to all ZTF data collected[3] by `Fink` from Nov/2019 to Mar/2021, for which an associated label was found in `SIMBAD`[4] or in Transient Name Server (`TNS`[5]). For each filter, the maximum flux was normalized to 1 and the time of first observation was shifted to zero. We then randomly sampled 100k AGN and 1 million non-AGNs alerts, and only considered objects with a minimum of 4 observed points in each filter. The resulting database contained 607772 alerts. Following [9], we regrouped objects into 2 larger categories, "AGN" and "non-AGN". The AGN category encloses [AGN, LINER, Blazar, BLLac, QSO][6]. Among `Fink-data`, 536621 alerts belong to the non-AGNs, while 71151 are AGNs. The resulting set was then equally divided into two samples: training and testing. Since each object can produce multiple alerts, we ensured that alerts from a given object were only present in one of the two samples. The `Fink-data` represents the state of the art of what can be done with real data.

Aiming to estimate the performance of our classifier in a LSST-like data environment, we put together a second database, hereafter, `ELAsTiCC_data` using data from the Extended LSST Astronomical Time-series Classification Challenge[7] (`ELAsTiCC`). It represents a more complex version of the

[2] `https://fink-broker.org/`
[3] `https://zenodo.org/record/5645609`
[4] `https://simbad.unistra.fr/simbad/`
[5] `https://www.wis-tns.org/`
[6] `SIMBAD` tags were used before the June 2022 taxonomy modification.
[7] `https://portal.nersc.gov/cfs/lsst/DESC_TD_PUBLIC/ELASTICC/`

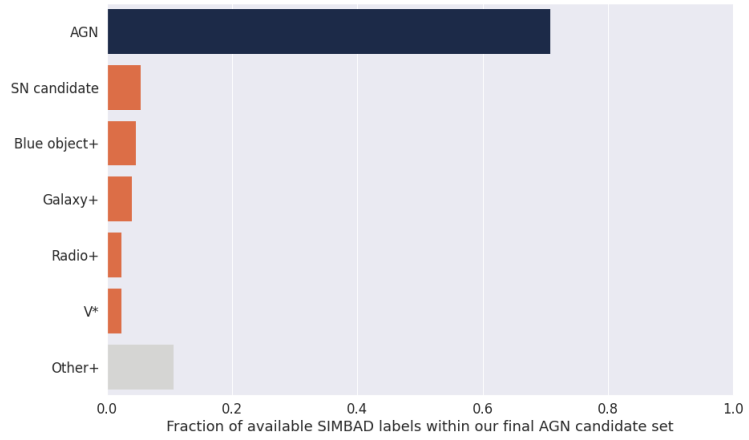

Figure 2: Broad categories of SIMBAD and TNS labels within the sample of identified AGN candidates. It follows SIMBAD taxonomy and the + marks classes which may include AGN sub-types.

Photometric LSST Astronomical Time-series Classification Challenge[8] (PLAsTiCC), held in 2018. Focusing in anticipating an LSST-like data and software environment, ELAsTiCC aims to test not only the classification power of broker systems, but also the resilience of their infrastructure and link with LSST facilities. It uses 32 different transient template models to generate simulated light curves using the Supernova Analysis (SNANA) [10] code. These are cut into alerts (each new photometric observation above detection threshold generate an alert) and daily streamed to broker teams. The brokers are expected to process the stream and send back their probability scores. Results will compare the performance of all teams who provided scores after an initial period of 3 months. A training sample composed of full light curves, and with a different cadence from the one used to create the test sample, was made available so broker teams could prepare/train their models for the challenge. This training sample is our starting point for the construction of the ELAsTiCC_data database. From it, we selected 50k AGNs and 50k non-AGNs objects with at least 4 points in 2 consecutive filters for training and another 50k AGN/50k non-AGN for testing.

3 Methodology

For each filter, we used the photometric points to compute the following features: maximum flux before normalization; standard deviation of the flux; number of points and mean signal to noise ratio. Moreover, we also added metadata to the features table, meaning: right ascension; declination; standard deviation and maximum absolute value of each color. Additionally, only for ELASTiCC_data, we added metadata information regarding: host galaxy redshift, host galaxy redshift error and host galaxy distance to the object.

The color calculation for this type of data is specially delicate. Since the sampling of the points is irregular, no pair of points exist at the exact same time in different filters, therefore we must interpolate each light curve so a proper colour estimation can be performed. In searching for a smooth function which would correctly capture the overall behavior of AGNs, we opted to use:

$$f(t) = \frac{1}{1 + e^{-At - e^{Bt} + C}} + D \tag{1}$$

This form was obtained by applying a traditional symbolic regression algorithm to noiseless simulated transient light curves via gplearn[9]. Once the code had converged and an analytical form propose, we manually substitute the float values by parameters, resulting in the reported functional form (Equation

[8] https://www.kaggle.com/c/PLAsTiCC-2018

[9] https://github.com/trevorstephens/gplearn

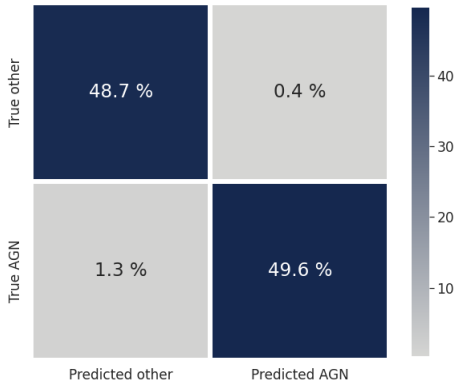

Figure 3: Confusion matrix results from the `ELAsTiCC_data` sample.

1). Therefore parameters A, B and C do not have any direct physical interpretation. However the parameter D was added latter on and represents the flux baseline. The minimization is performed using the least squared function from the python module iMinuit ([11]). Figure 1 shows an example of a real alert from `Fink-data`, alongside the estimated fluxes. Color was computed by subtracting flux values between two consecutive filters, [g-r] for `Fink-data` and [u-g, g-r, r-i, i-z, z-Y] for `ELAsTiCC_data`, at the condition that both concerned filters contains at least 4 observed points (considering forced photometry but not upper limits).

Once the feature matrix was constructed, classification proceeded using the random forest algorithm from the python module scikit-learn ([12]). This method uses an ensemble of decision trees, each built from a different sub-sample of the data. When new data is inputted, the answer from each tree is counted as a vote, and the probability outputted by the forest comes from the proportion of each vote. In what follows, all results were obtained using 100 trees. Conducted attempts using a gradient boosted tree algorithm, as well as attempts to increase the number of trees, lead to no further improvements in the classification score.

For `Fink-data`, an optimized training sample was built by using an active learning strategy based on uncertainty sampling [13]. We started the learning cycle with 5 AGN and 5 non-AGNs randomly selected from the sample of available labels, and trained a random forest using these 10 objects while the rest of the data was used for testing. The alert in the testing sample for which the random forest is the most uncertain about (probability closest to 50% percent) is moved from testing sample to training sample in each cycle. We performed 2000 loops, resulting in a final training sample of 2010 objects. The entire process was performed 10 times in order to briefly access the impact of different initial conditions. In the case of `ELAsTiCC_data`, representativeness was ensured by construction, thus no active learning loop was required.

4 Results and conclusion

The active learning process will, by construction, build a training data set approximately made of 50% AGNs and 50% non-AGNs. In the case of unbalanced data set such as `Fink-data`, this property is crucial. A balanced and informative training set ensures that the random forest is learning to separate types of alert, and not just learning the statistical distribution of this particular dataset. Furthermore a balanced training dataset guarantees the classifier optimal threshold to be around 0.5.

When applied to all the non queried objects, the classifiers built with active learning achieved on average 98.6 % accuracy, 95.9 % precision and 91.9 % recall on finding AGNs, with a standard deviation of 0.1%, 0.7% and 0.8% respectively.

We conclude that the uncertainty sampling was efficient at building a training sample. It also ensures that the classifier will perform well when used on another dataset with different statistical properties. Among the 10 runs, we kept the classifier with highest product recall $\times$ score, which in our case also satisfies the F1 criteria. Applying this classifier to the testing sample (not involved in training), we achieve a score of 98.0 % accuracy with 93.8 % precision and 88.5 % recall on AGNs. Finally the

classifier has been tested directly on the Fink alert stream. All alerts between Jan/2019 to Jan/2020 were processed and we evaluate the results for every labelled alerts (excluding objects used in the training sample). Figure 2 shows the distribution of labels within the photometrically predicted AGNs. Results from the `ELAsTiCC_data` sample are shown in Figure 3. Feature importance analysis on results obtained from the `Fink-data` shows that all features play a similar role in enabling classification. For the `ELAsTiCC_data`, the number of points acts like a first layer of classification. This was expected, since this data set contain a high percentage of transient classes. These are objects whose brightness are only significant during a limited time window, which makes them more easily distinguishable from AGNs, that are persistent sources.

Both models described in this work are now integrated to the `Fink` broker, lively processing alerts from ZTF and `ELAsTiCC`. The filtered ZTF stream will be directed to the AGN community interested in spectroscopic follow-up of individual events. We foresee the development of similar models focused on sub-types of AGNs chosen by `Fink` users. Finally, we intend to use the AGN scores to further filter out the data given to the supernova classifiers in `Fink`, thus allowing for an even higher precision in the estimates delivered by the broker.

All the work presented in this paper is publicly available on GitHub [10].

Acknowledgements

We thank David O. Jones for making simulations available for this project. This work was developed within the `Fink` community and made use of the `Fink` community broker resources. `Fink` is supported by LSST-France and CNRS/IN2P3.

Impact statement

Although the results presented in this paper are based on ZTF data or simulations, the real goal of `Fink` is to prepare an appropriate environment to deal with LSST data. The experience and results presented here will guide the development of more efficient pipelines, which optimize the use of manpower and spectroscopic follow-up facility as well as computer resources. Given the data volume and complexity to be delivered by LSST, the development of efficient photometric classifiers is the main bottle neck to be solved in order to ensure we will be able to fully exploit data which took so much effort to acquire.

References

[1] P. Padovani, D. M. Alexander, R. J. Assef, and et al. Active galactic nuclei: what's in a name? *The Astronomy and Astrophysics Review*, 25(1):2, August 2017. doi: 10.1007/s00159-017-0102-9.

[2] Hirofumi Noda and Chris Done. Explaining changing-look AGN with state transition triggered by rapid mass accretion rate drop. *Monthly Notices of the Royal Astronomical Society*, 480(3): 3898–3906, November 2018. doi: 10.1093/mnras/sty2032.

[3] P. Sánchez-Sáez, H. Lira, L. Martí, and et al. Searching for Changing-state AGNs in Massive Data Sets. I. Applying Deep Learning and Anomaly-detection Techniques to Find AGNs with Anomalous Variability Behaviors. *The Astronomical Journal*, 162(5):206, November 2021. doi: 10.3847/1538-3881/ac1426.

[4] Benny Trakhtenbrot, Iair Arcavi, Claudio Ricci, and et al. A new class of flares from accreting supermassive black holes. *Nature Astronomy*, 3:242–250, January 2019. doi: 10.1038/s41550-018-0661-3.

[5] V. A. Masoura, G. Mountrichas, I. Georgantopoulos, and et al. Disentangling the AGN and star formation connection using XMM-Newton. *Astronomy and Astrophysics*, 618:A31, October 2018. doi: 10.1051/0004-6361/201833397.

[6] Mary Loli Martínez-Aldama, Bożena Czerny, Damian Kawka, and et al. Can Reverberation-measured Quasars Be Used for Cosmology? *The Astrophysical Journal*, 883(2):170, October 2019. doi: 10.3847/1538-4357/ab3728.

[10] `https://github.com/astrolabsoftware/fink-science/tree/master/fink_science/agn`

[7] Maria T. Patterson, Eric C. Bellm, Ben Rusholme, and et al. The zwicky transient facility alert distribution system. *Publications of the Astronomical Society of the Pacific*, 131(995):018001, nov 2018. doi: 10.1088/1538-3873/aae904. URL `https://doi.org/10.1088/1538-3873/aae904`.

[8] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[9] M. Leoni, E. E. O. Ishida, J. Peloton, and A. Möller. Fink: Early supernovae ia classification using active learning. *Astronomy & Astrophysics*, 663:A13, jul 2022. doi: 10.1051/0004-6361/202142715. URL `https://doi.org/10.1051%2F0004-6361%2F202142715`.

[10] Richard Kessler, Joseph P. Bernstein, David Cinabro, Benjamin Dilday, Joshua A. Frieman, Saurabh Jha, Stephen Kuhlmann, Gajus Miknaitis, Masao Sako, Matt Taylor, and Jake Vanderplas. SNANA: A Public Software Package for Supernova Analysis. *Publications of the Astronomical Society of the Pacific*, 121(883):1028, September 2009. doi: 10.1086/605984.

[11] Hans Dembinski and Piti Ongmongkolkul et al. scikit-hep/iminuit. Dec 2020. doi: 10.5281/zenodo.3949207. URL `https://doi.org/10.5281/zenodo.3949207`.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, and et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[13] E. E. O. Ishida, R. Beck, S. González-Gaitán, and et al. Optimizing spectroscopic follow-up strategies for supernova photometric classification with active learning. *MNRAS*, 483(1):2–18, February 2019. doi: 10.1093/mnras/sty3015.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] This is added to our github repo[11]10
- Did you include the license to the code and datasets? [Yes] on github
- Did you include the license to the code and datasets? [No] already made public by others

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [No]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes]

(b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [No]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]