Correlation between gamma-ray sources and high energy neutrinos using deep learning

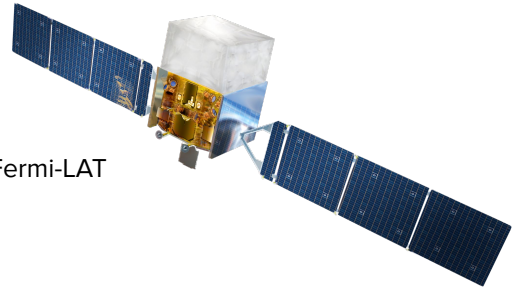
Enzo Oukacha

Supervisor : Yvonne Becherini

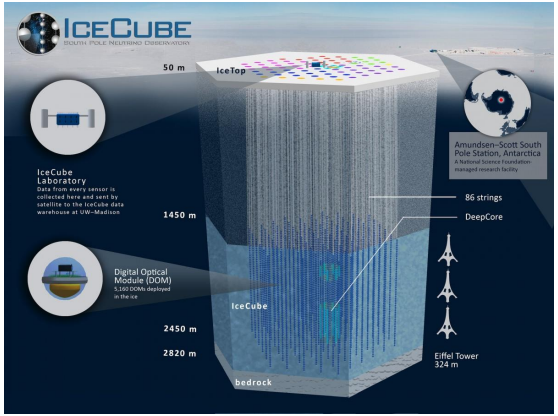
APC[1]

1. Laboratoire Astroparticule et Cosmologie

Motivation of CoNIC

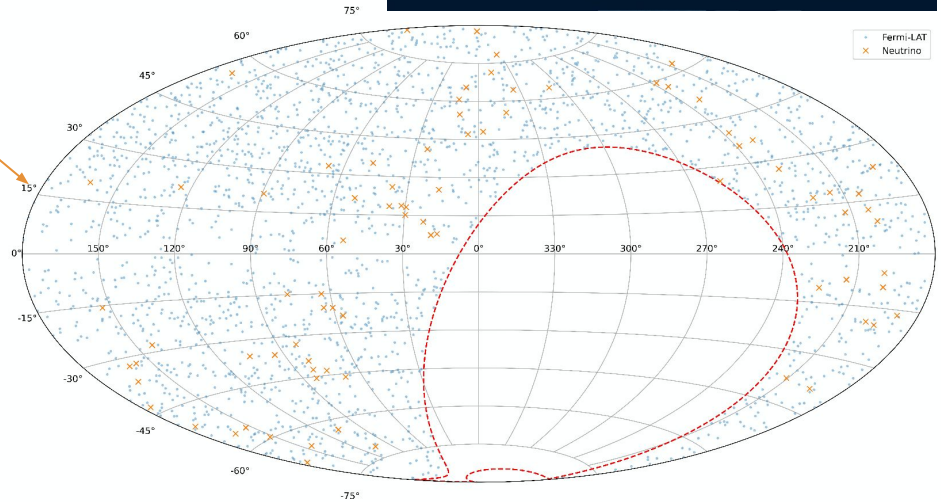


Fermi-LAT



Motivation:

No clear association between the high energy neutrinos detected by the IceCube experiment (alerts) in the South Pole and the extragalactic gamma-ray sources (Active Galactic Nuclei) seen by the Fermi-LAT satellite



Goal of CoNIC:

Find an indirect association between neutrinos and gamma ray sources with Deep learning methods.

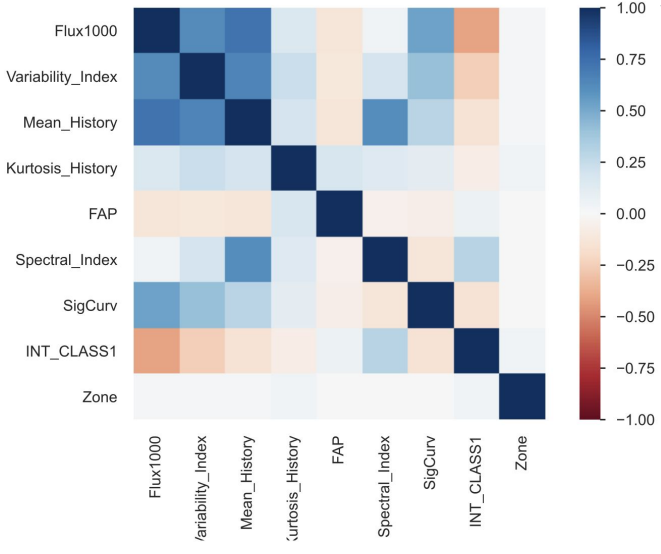
Supervised learning approach

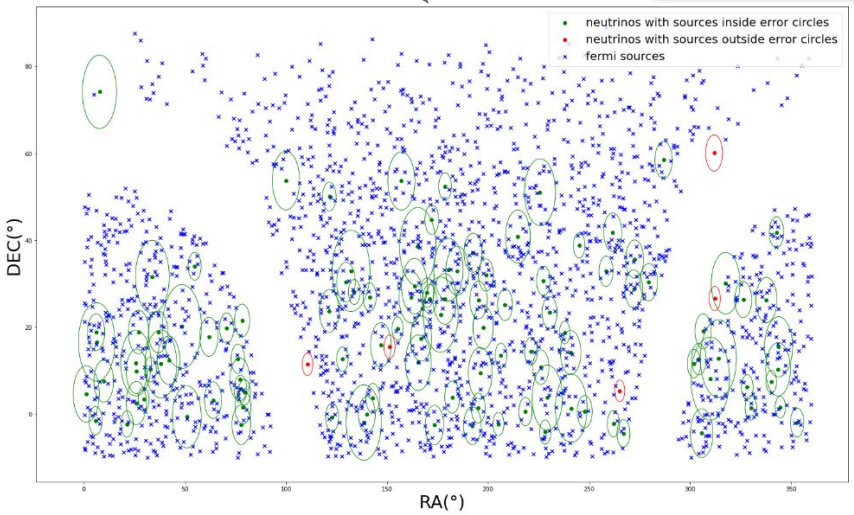
Definition of ON zones (labels)

- **Selection of quality sources in 4FGL-DR4**
- **Selection of quality events in IceCat-1**

Selection of Fermi sources features

Supervised approach with tabular data



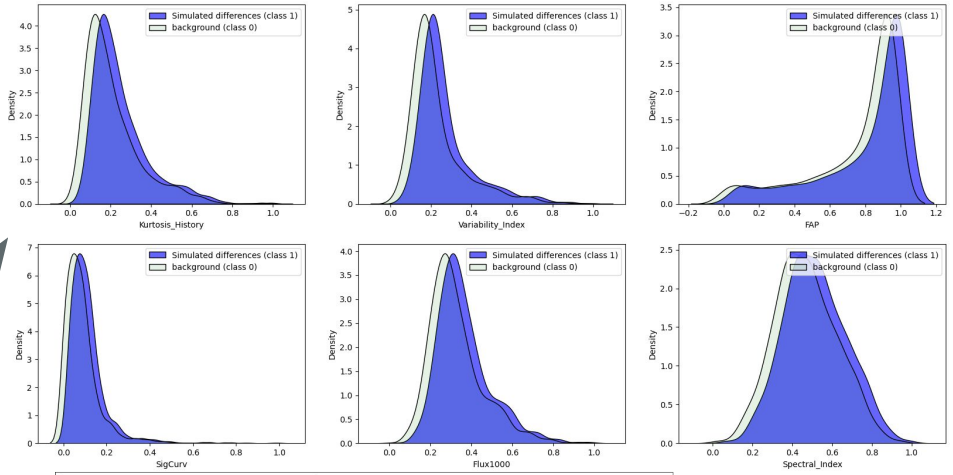


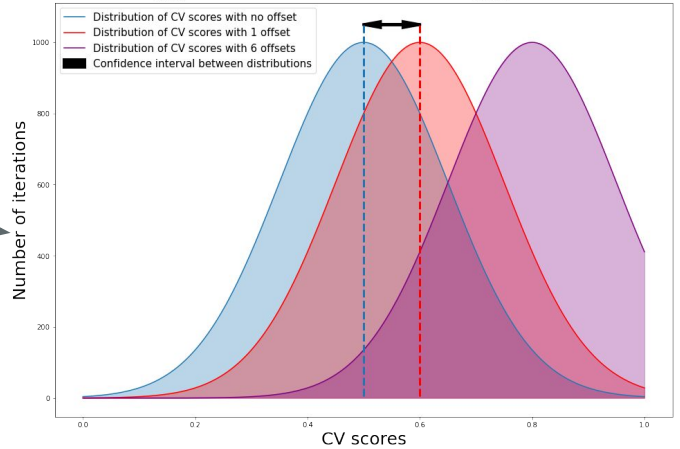
Characterization of the background

Simulation of differences :

1) Select all background data in the original dataset (class 0).
2) Duplicate this dataset and merge the duplicated dataset (target 1) with the original one (target 0)
3) Train the NN/Classifier and get a cross validation score distribution.
4) Shift the distributions of 1 or multiple features for class 1 sources and check results.

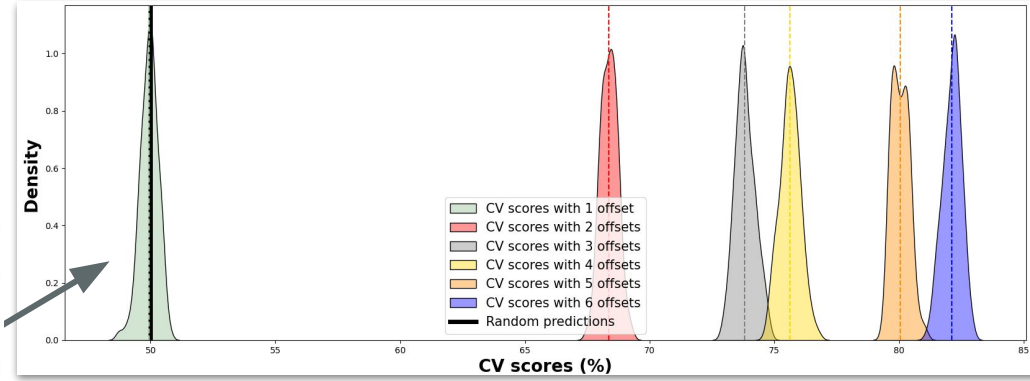
Hypothetic CV distributions

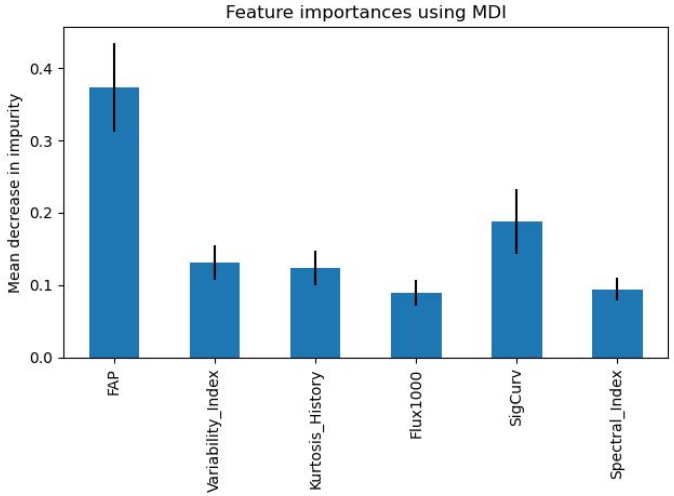




Sensitivity of the Random Forest

Classifier	Training conditions	CV score
RF	7f-7d	87% (+/- 1.61%)
RF	6f-6d	85% (+/- 2.27%)
RF	6f + random gauss	85% (+/- 1.19%)
RF	6f + correl feature	95% (+/- 1.32%)
RF	umap 2d embeddings	62% (+/- 3.50%)





Complementary approach based on Time-Series

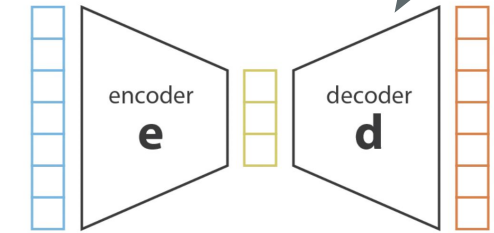
Light curves (time-series) generated with Fermipy[2]

Supervised approach

Unsupervised/ Self-supervised approach

Encoding of the data and visualization

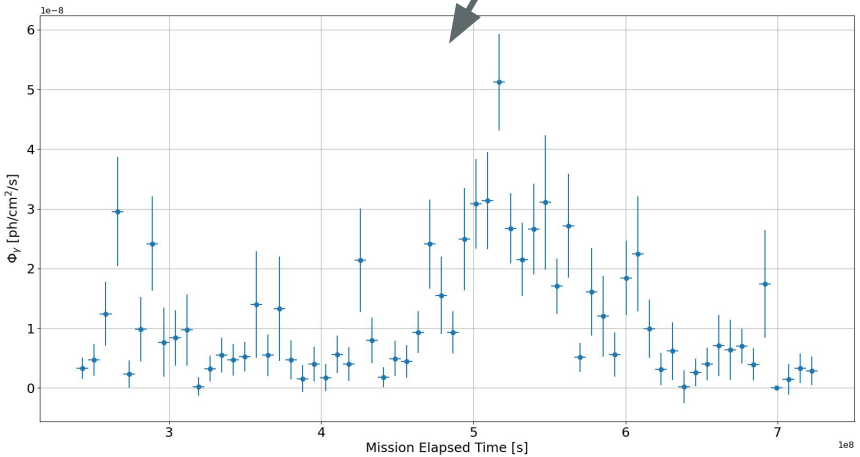
Down-stream task



x — initial data in space R^n

e(x) — encoded data in latent space R^m (with m<n)

d(e(x)) — encoded-decoded data back in the initial space R^n



2. Wood, M., Caputo, R., Charles, E. et al. 2017 PoS

Efficiency of the method

Problem(s) with the supervised approach :

- Lack of data will end up in bad performances
- Need numerous good quality labels for meaningful results
- Task-specific models may be good for one task but bad if we need to generalize/extend to more global problems

Problem(s) with the unsupervised approach :

- Lack of data will end up in a badly structured latent space.
- Interpretation of the results may be difficult if not enough attention has been dedicated to the representation of the data

Adapted pipeline for Time-Series analysis

Simulated TS for training/ validation purpose

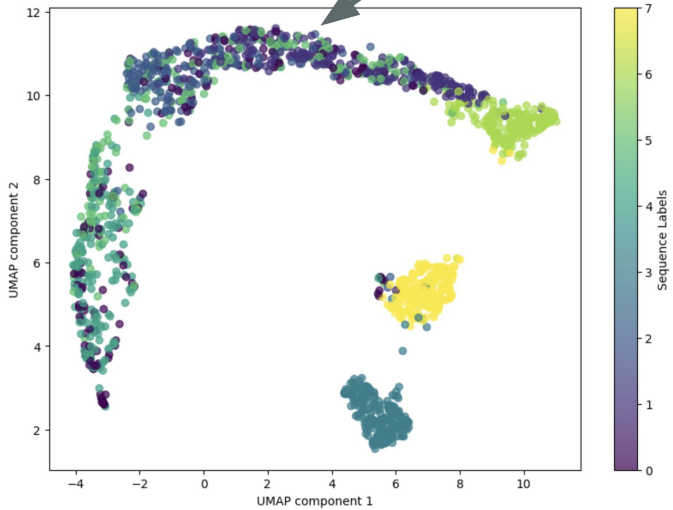
Unsupervised/ Self-supervised approach

Observational TS for testing purpose

Pre-trained encoder based model

Representation of the data in a reduced dim space

Downstream task based on Transfer learning



Incoming work

On-going work :

- Baselines obtained with a supervised classification approach based on a transformer's/cnn model.
- Implementation of a backbone model for the whole pre-training task.
- Use of different auto-encoders models (standard/variational/adversarial) to get the best structured latent space possible with the simulated data.
- Transfer-learning for the test phase with observational data.