UFJF CYGNO ML Projects

Rafael A. Nóbrega

with Igor Pains, Guilherme Lopes, Jórdan Leite and Luan Gomes

Team and Projects on ML

- **Team**
  - Rafael A Nóbrega → Prof. at UFJF
  - Guilherme Lopes → PhD student
  - Igor Pains → PhD student
  - Jordan Leite → PhD student
  - Luan Gomes → Undergraduate student (soon in our Master's program)

- **Algorithms under implementation**
  - U-Net to improve SNR of images (possible extensions)
  - CNN for trigger purposes
  - ANN, Gradient Boosting, Random Forest and CNN for NR detection

ML past projects on CYGNO

- Clustering - iDBSCAN
    - *JINST journal (2020) and Abritta's PhD thesis (2020)*
- Filtering impact on CYGNO data applied to LEMON *(using standard filters and U-Net)*
    - *Guilherme's Master thesis (2021)*
- Long track clustering - Directional DBSCAN
    - MEASUREMENT SCIENCE and TECHNOLOGY (2023) *and Pains' Master thesis (2024)*

Trigger

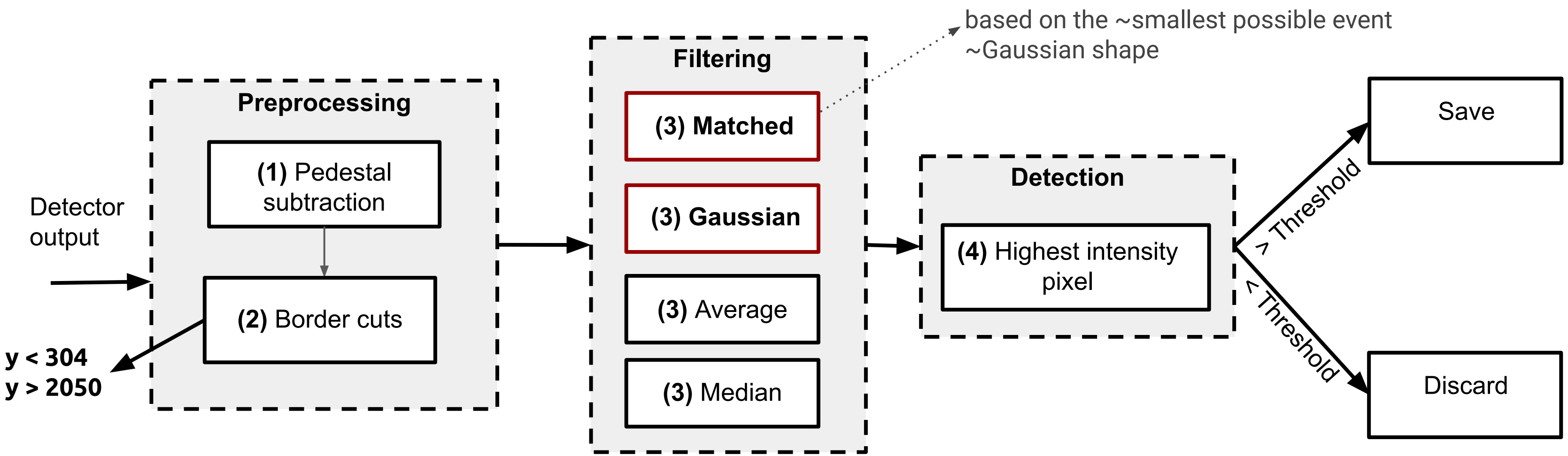
- Main goal → create and study “fast” algorithms that can select images with signal (or discard noise-only images).

What was done…

- **Two approaches were considered:**
  - Based on **filtering.**
  - Based on **CNN.**

- **Comparative analysis:**
  - **Detection performance** on low energy simulated signals.
  - **Time** analysis.
  - Comparison with the **reconstruction algorithm.**

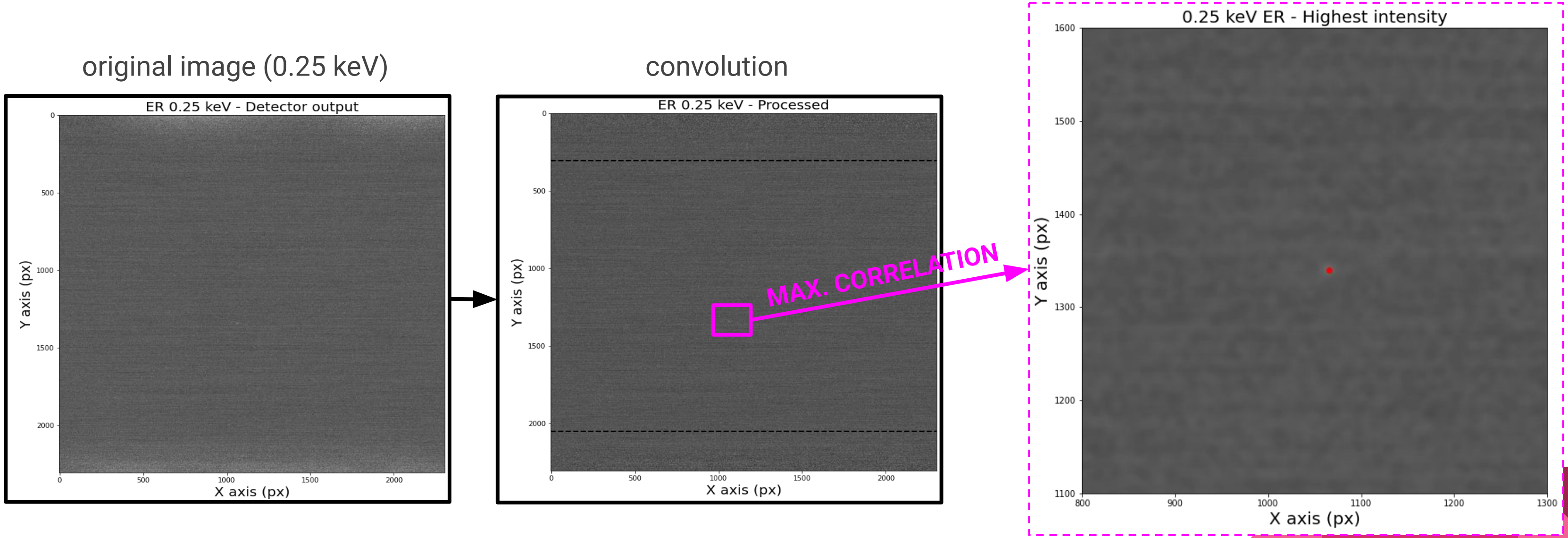
Trigger based on filtering (correlation)



Filter parameters and detection threshold selected based on training data

Trigger based on filtering (correlation)

Example…

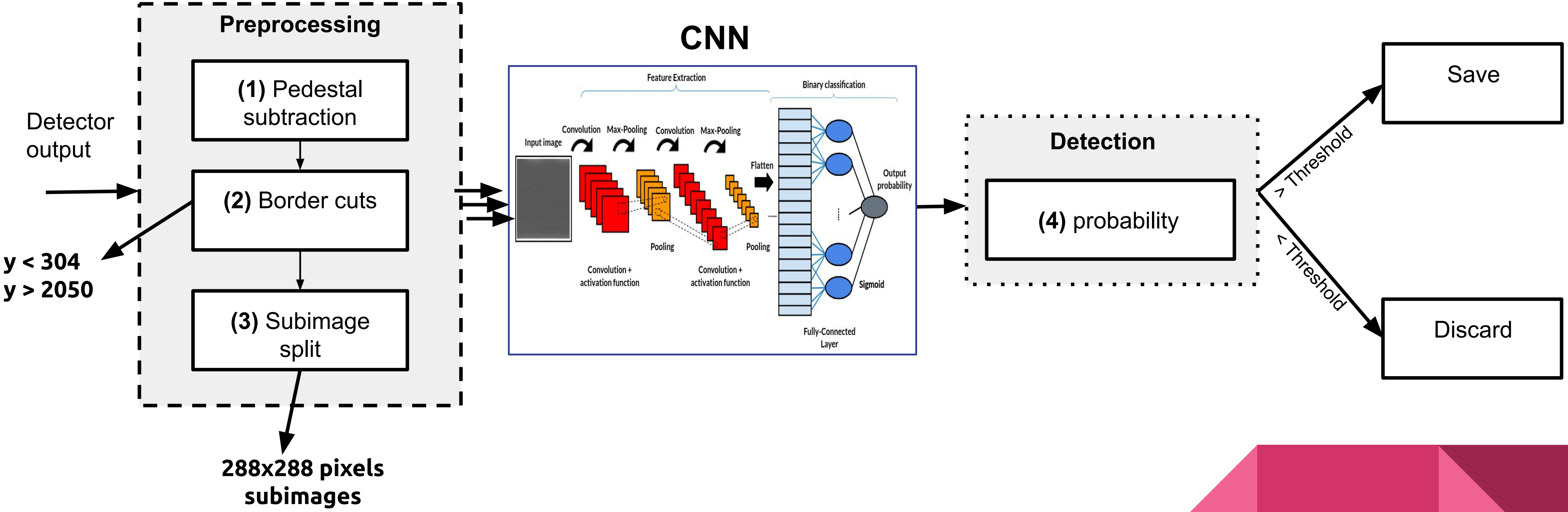


original image (0.25 keV)

convolution

detection (thresholding)

Trigger based on CNN



Detector output

Preprocessing

(1) Pedestal subtraction

(2) Border cuts

(3) Subimage split

y < 304
y > 2050

288x288 pixels subimages

CNN

Detection

(4) probability

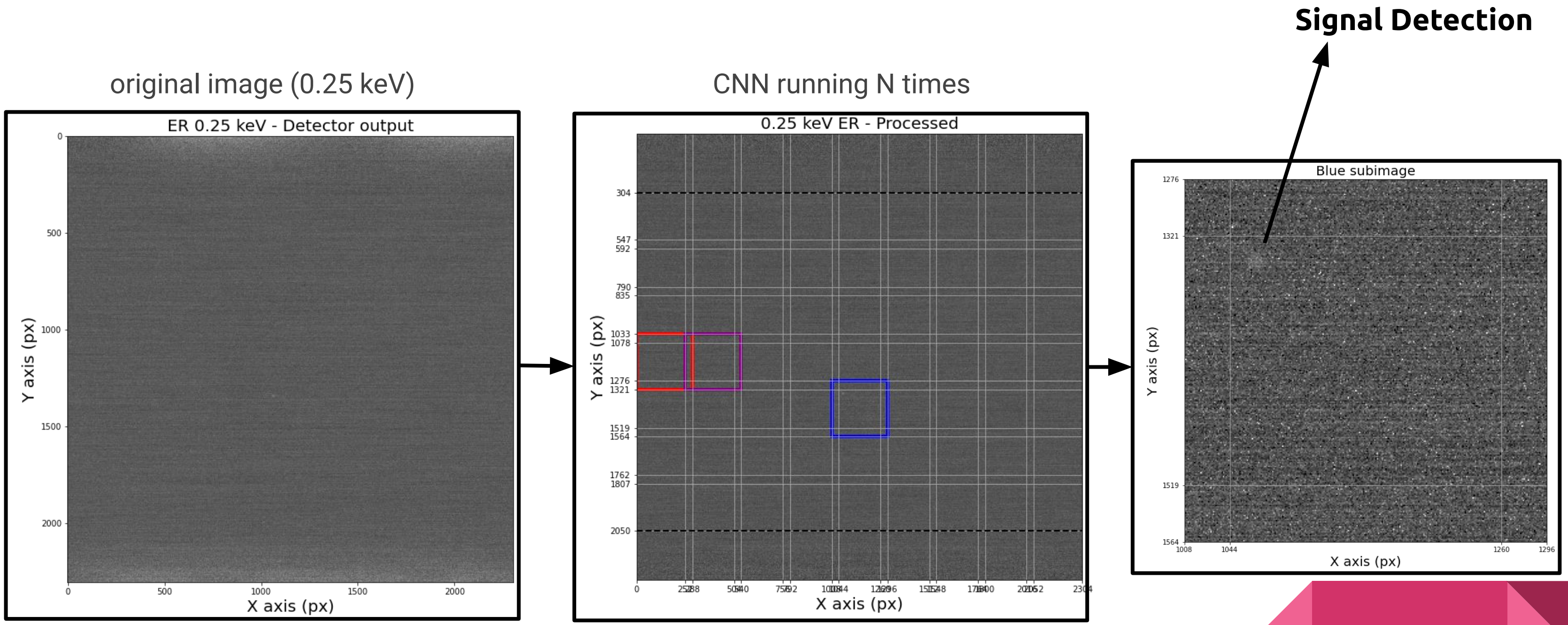
> Threshold

< Threshold

Save

Discard

Trigger based on CNN



original image (0.25 keV)

CNN running N times

Signal Detection

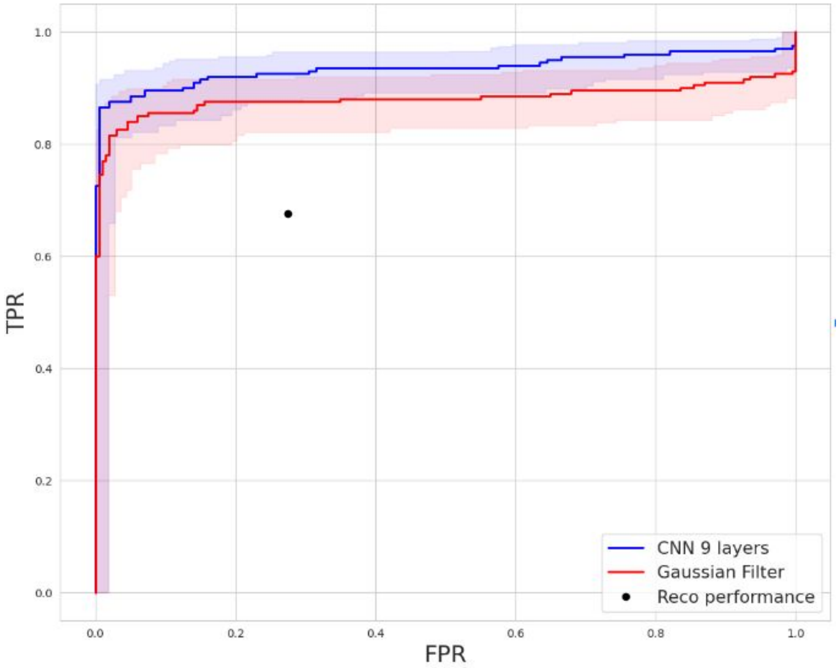
Dataset

- **Training:**
  - **Noise dataset:** 600 images from pedestal runs (Run 4 underground).
  - **ER and NR signal simulation:** 600 images each containing **0.25-1 keV** signals added to pedestal runs (different from noise dataset).
- **Validation:**
  - **Noise dataset:** 200 images from pedestal runs.
  - **ER and NR signal simulation:** 200 images each containing **0.25-1 keV** signals.
- **Test:**
  - Same configuration as validation.

Trigger performance

- **CNN**
    - **80% detection eff.** *(fixed)*
    - **0.5% false alarm.**
- **Gaussian filter**
    - **80% detection eff.** *(fixed)*
    - **10% false alarm**
- Both methods **outperforms** the **reconstruction** code in detecting **0.25 keV** signals.
- All methods can **easily detect** signals with energies above **0.5 keV.**

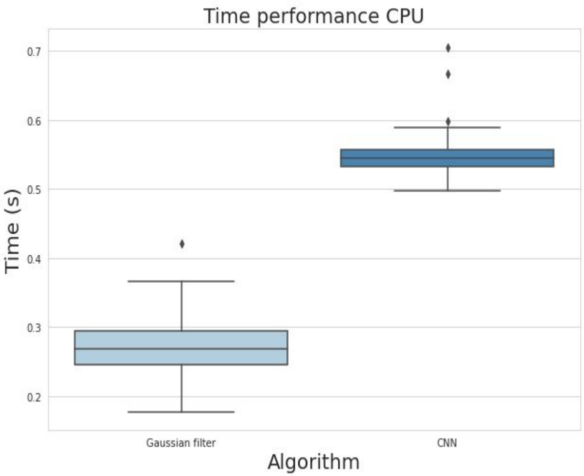
Test for 0.25 keV

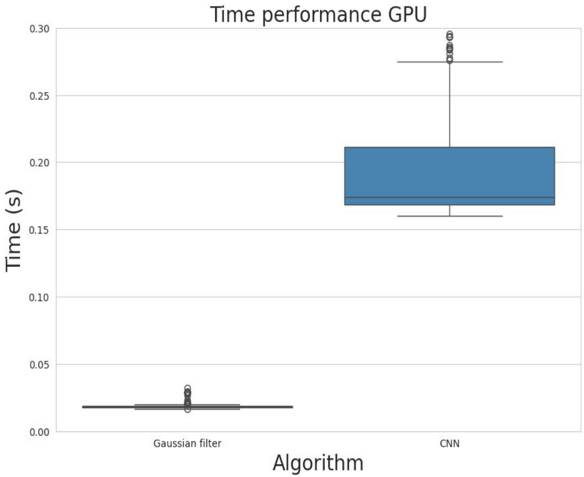


Processing time

- **CNN**
  - **CPU → ~0.55 sec**
  - **GPU → ~0.2 sec.**

- **Gaussian filter**
  - **CPU → ~0.25 sec**
  - **GPU → ~0.02 sec.**





Higher detection performance comes at the cost of slower processing time.

Next steps

- Study methods to simplify the CNN model (on going)
  - Bit reduction, weight combination, pruning and vectorization.

time reduction

- Test popular CNN architectures such as AlexNET, GoogleLeNet, Unet with necessary adaptations.

performance improvement

- Apply the CNN on the DAQ machine.

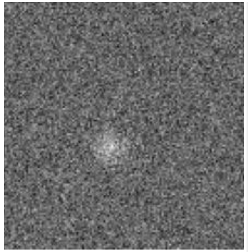
- Write a paper based on these results.

U-Net application

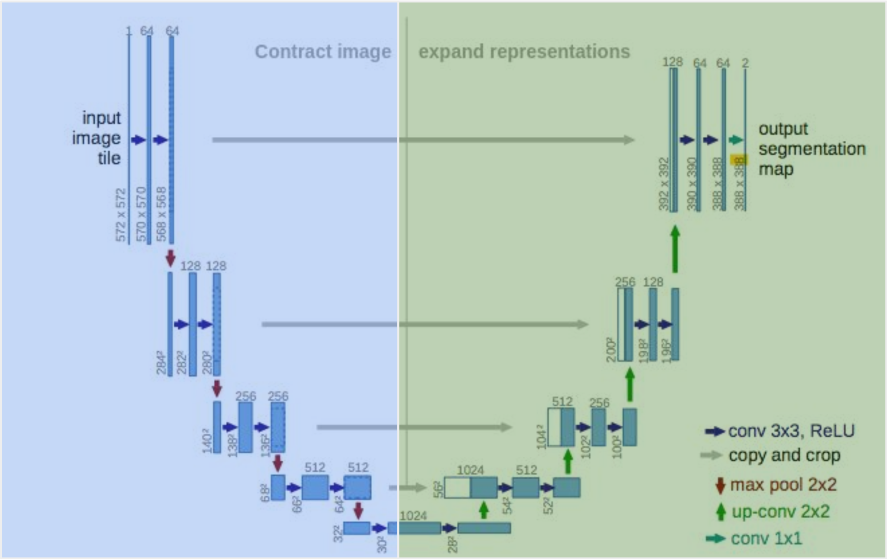
- Main goals:
  - improve the pre-processing stage (semantic segmentation)
    - improve clustering performance
    - reduce clustering time
    - Improve detection, energy estimation, etc

  - investigate performance for:
    - *clustering and particle identification*

U-Net pixel-wise classification



Simulation Image

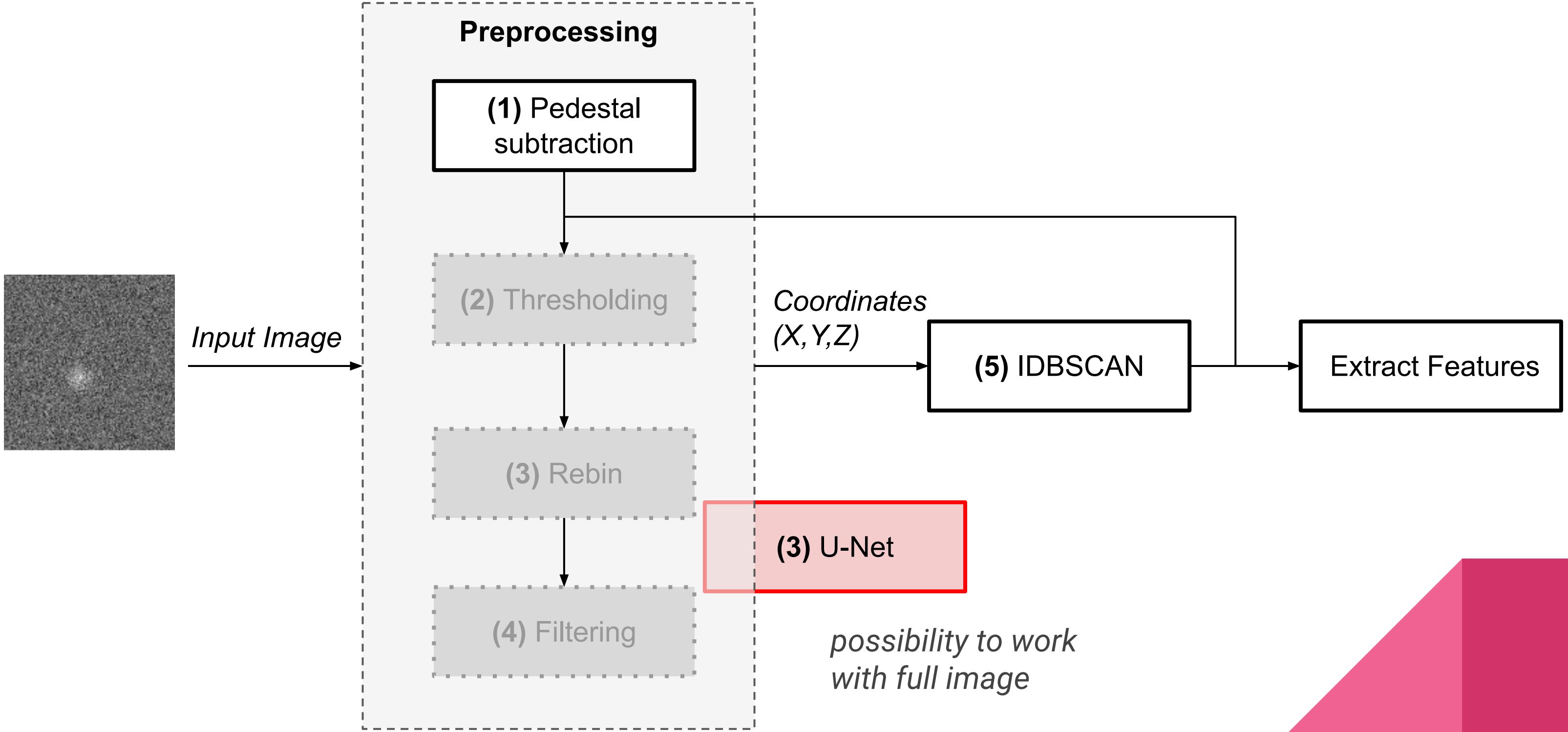


Truth Image

Contract step is used to build feature maps using convolution;

Expand step will upsample these features to allow pixel wise-relationship between input and output;

Using U-Net in reconstruction algorithm



Input Image

Preprocessing

(1) Pedestal subtraction

(2) Thresholding

(3) Rebin

(4) Filtering

(3) U-Net

Coordinates (X,Y,Z)

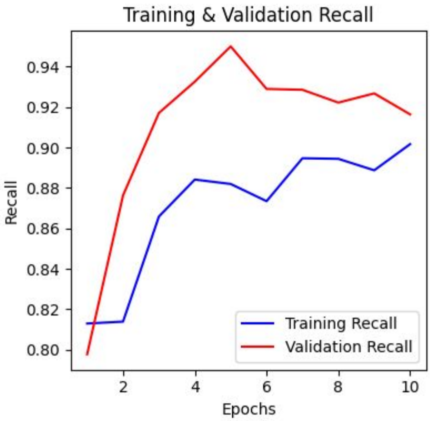
(5) IDBSCAN

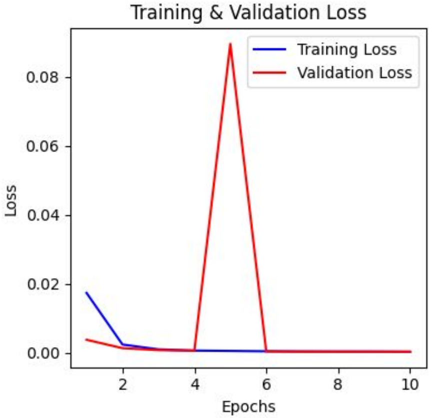
Extract Features

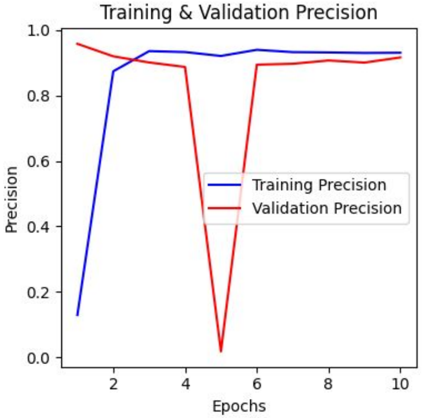
possibility to work with full image

Training U-Net

- **Focal Loss** used to deal with Imbalanced data (most part of image is background)
- **Precision** and **recall** to evaluate performance of pixel-wise classification on 0.5 threshold.

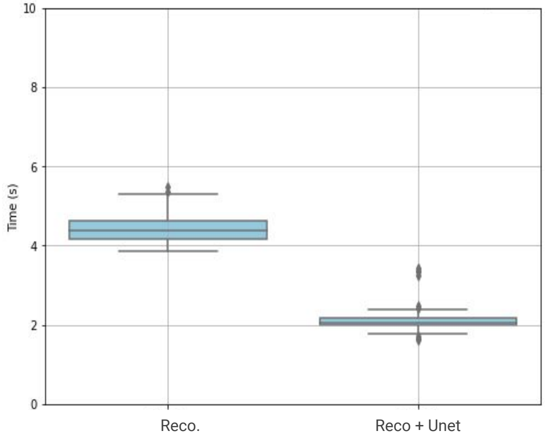

Training & Validation Recall


Training & Validation Loss


Training & Validation Precision

Preliminary results of U-Net *(0.25 and 0.50 KeV)*

Improvements on processing time

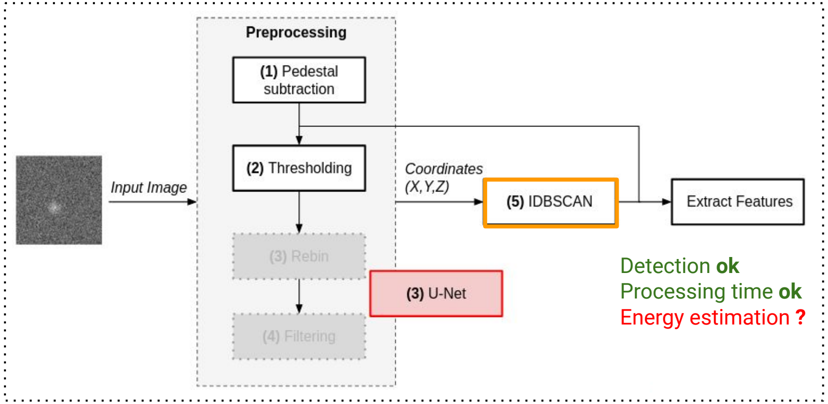


Improvements of event detection

Method	Reco	Unet + Reco
Detected events (%)	**68**	**86**

Next step

- Optimizes DBSCAN parameters for U-Net outputs
- Study impact of U-Net on energy estimation (on going)
- Write a paper based on these results.



Detection **ok**
Processing time **ok**
Energy estimation **?**

Future possibilities with U-Net

- Use **transfer learning** approach to explore new possibilities:
    - energy estimation
    - directionality detection
    - particle classification

NR detection

- Main goal → detection of NR events

What was done and very next steps

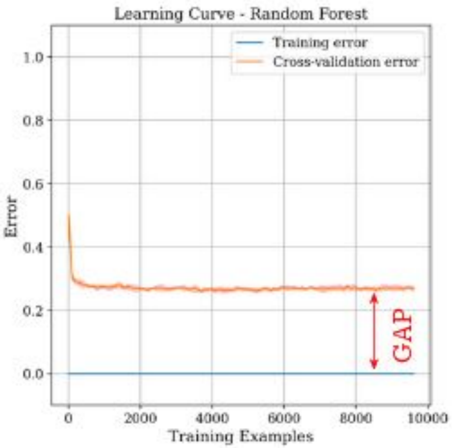
- Find a suitable architecture (using all the 36 variables):
  - ANN
  - Gradient Boosting (GBC)
  - Random Forest (RFC)
  - CNN

recently started
very preliminary

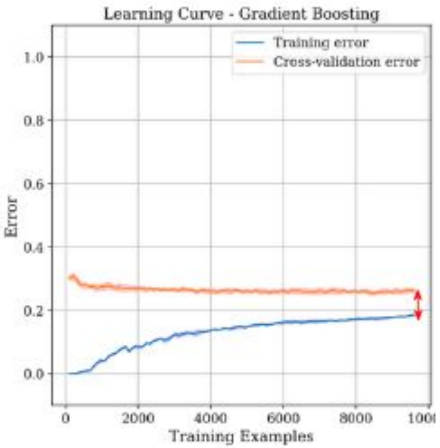
- Feature selection
  - with Random Forest and Gradient Boosting (not applied yet)
  - Genetic Algorithm (on going)
- Apply Atul's work inputs and outcomes (on going)
  - Use selected variables
  - Use dataset (simulated and real events)
  - Apply same pre-processing

Training data with RFC and GBC

Random Forest Classifier

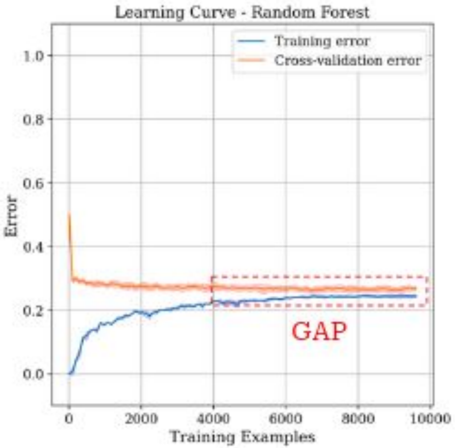


Gradient Boosting Classifier

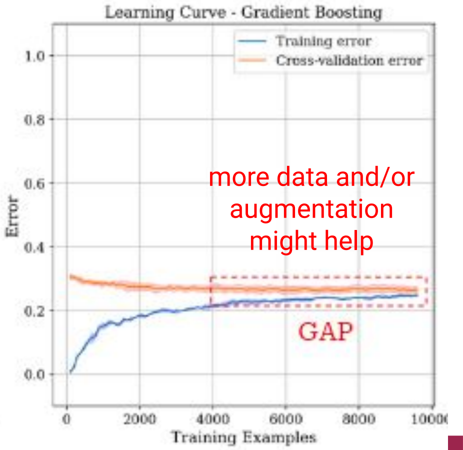


After HalvingGridSearchCV
for best hyperparameters

Random Forest Classifier



Gradient Boosting Classifier



more data and/or
augmentation
might help

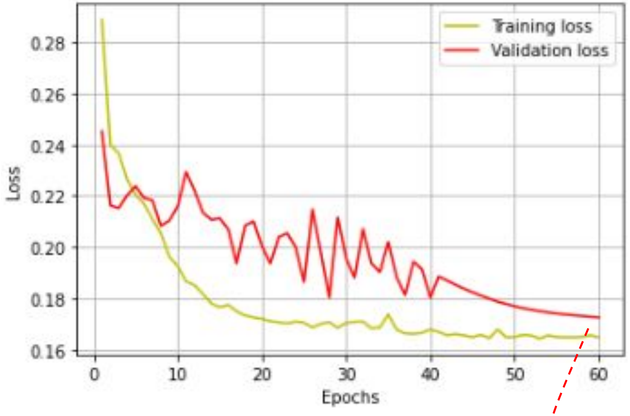
Used dataset

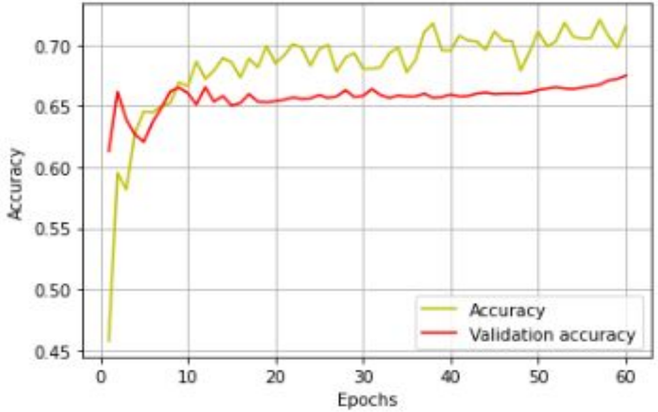
Energies in a range from **1-60 keV**
6 energy levels
1000 events per level

Training data with ANN

Used dataset

Energies in a range from **1-60 keV**
6 energy levels
1000 events per level

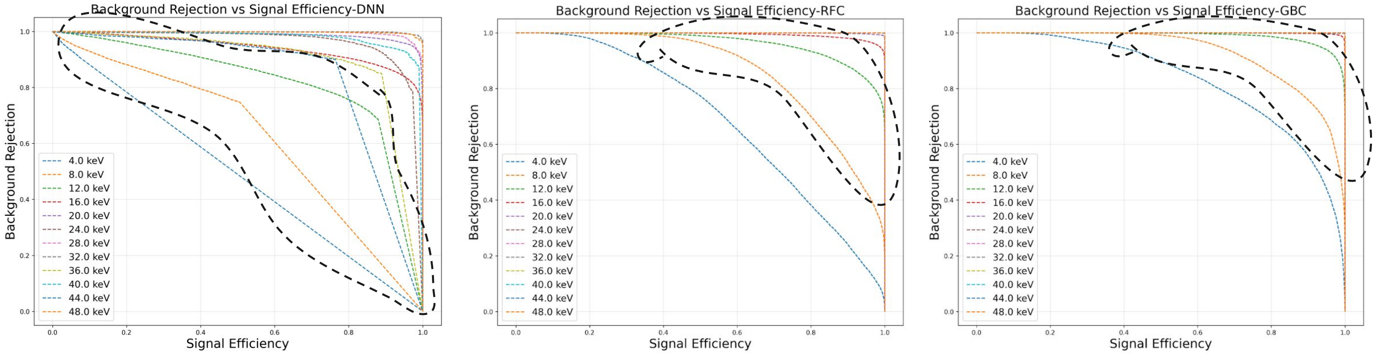




Could have continued…

Very Preliminary Results

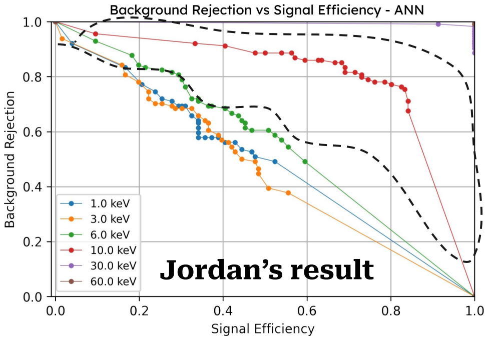
Training with all energies - Background Rejection vs Signal Efficiency



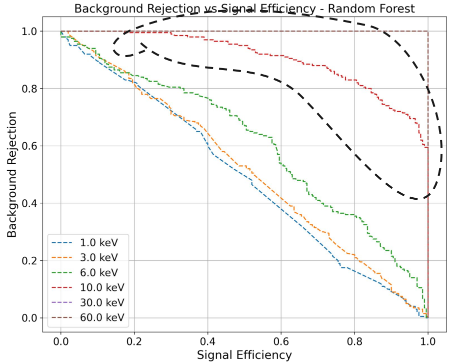
From Atul's thesis
Figure 6.19

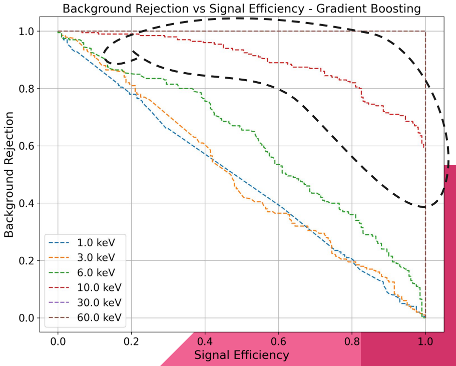
Comparing

8 / 12 KeV
to
10 keV



Jordan's result





Conclusions and next steps

- The first step has been taken toward continuing the Atul’s work
- A very preliminary study has been done for ER/NR classification
  - RFC, GBC, ANN (CNN to be included)
- Results showed divergences and similarities with Atul’s results
  - By now, the comparison is not trivial and direct

Atul’s dataset

Energies in a range from **2-50 keV**
25 energy levels
~10000 events per level

Used dataset

Energies in a range from **1-60 keV**
6 energy levels
1000 events per level

- **Next steps:**
  - Pre-processing (Geo and noise cut, outlier removal)
  - Robust feature selection
  - Data augmentation
  - Apply Atul’s work inputs and outcomes
  - Include other background events (and possibility PMTs)

Re-evaluate performance

General Conclusions

General Conclusions

- By the moment we are working on three fronts:
  - Image Trigger by CNN *(could/should be merged with the PMT signal)*
    - *Igor Pains*
  - SNR improvement using U-Net *(that can be extended to achieve other goals: clustering, classification, …)*
    - *Guilherme*
  - NR detection using ANN, RF, GBC, CNN
    - *Jórdan and Luan*

- NR detection has just started → many things to do

- All of them can be further studied and optimized (on going)