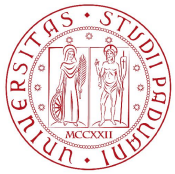


Physics of Data

MAGIC Data Analysis

Laboratory of Computational Physics - Mod. A

Lorenzo Cavezza
Giulia Doda
Giacomo Longaroni
Laura Ravagnani

Outline

Features Analysis

Outline

Features Analysis

Data Classification

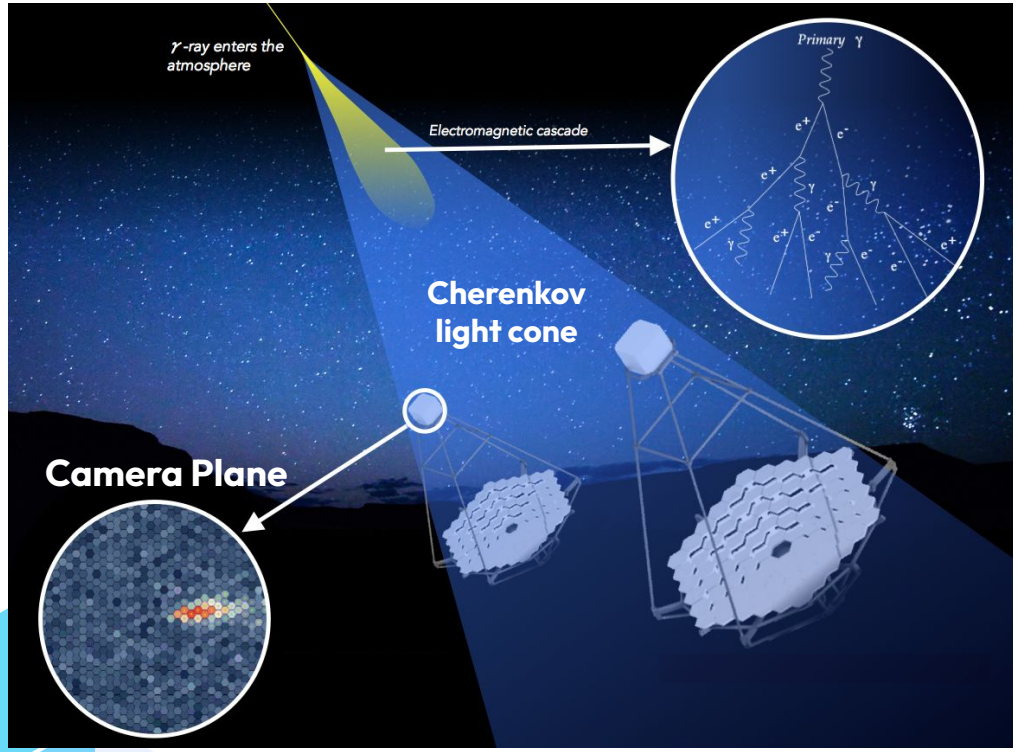
Outline

Features Analysis

Data Classification

Collection Time

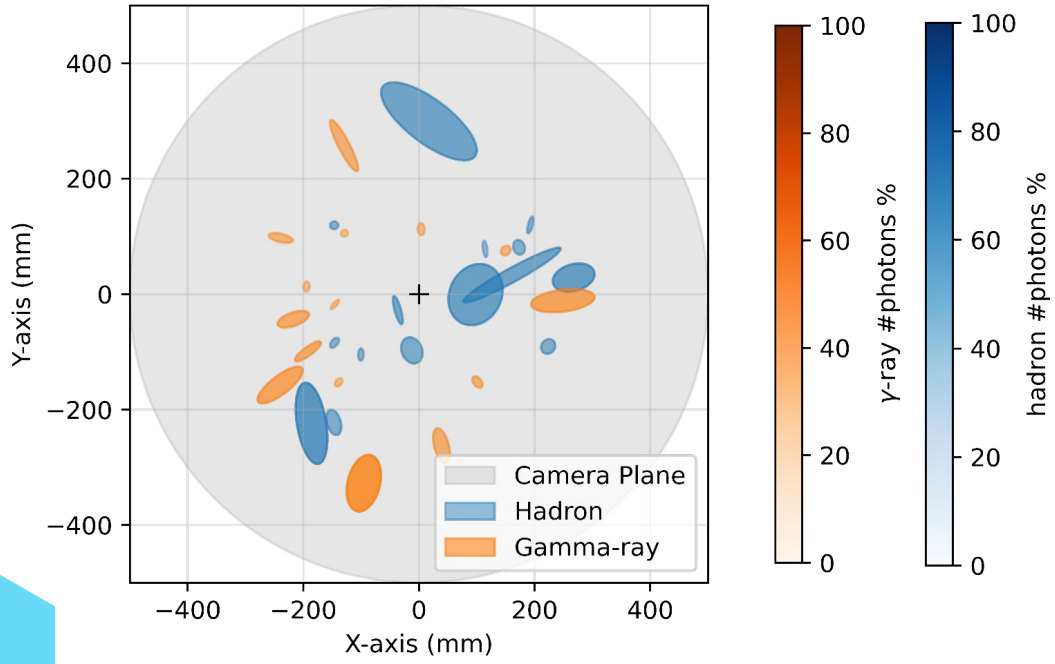
Dataset



- **MAGIC**: two gamma-ray imaging atmospheric Cherenkov telescopes
- **Monte Carlo simulated data**: max 50 GeV for primary gamma ray
- **Dataset dimension**: 19020 samples
  10 features
- **Dataset asymmetry**: 65% signal
  35% bkg

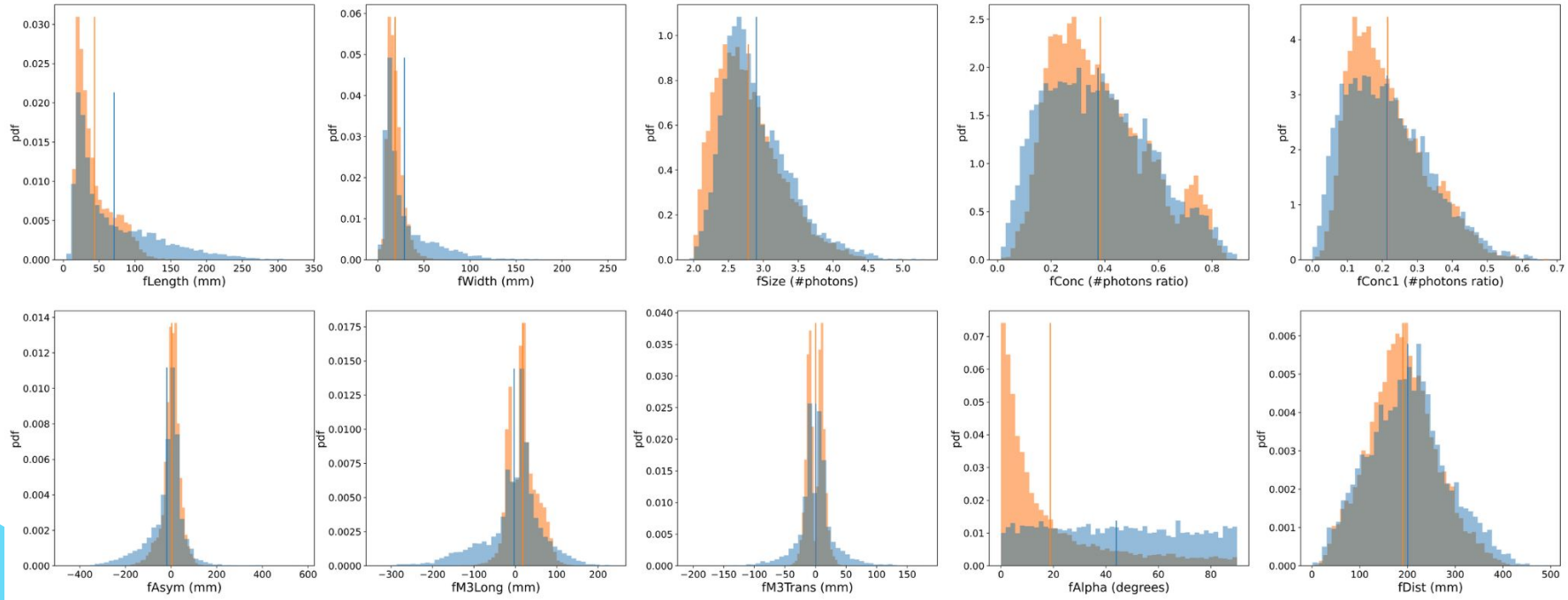
Data Features	
fLength (mm)	fAsym (mm)
fWidth (mm)	fM3Long, fM3Trans (mm)
fSize (#photons)	fAlpha (deg)
fConc, fConc1	fDist (mm)

Image reconstruction



- Images of the showers are **elongated clusters**
- **Ellipses orientation**
  - **Signal**: radially
  - **Background**: random
- Each shower is composed by a different amount of photons

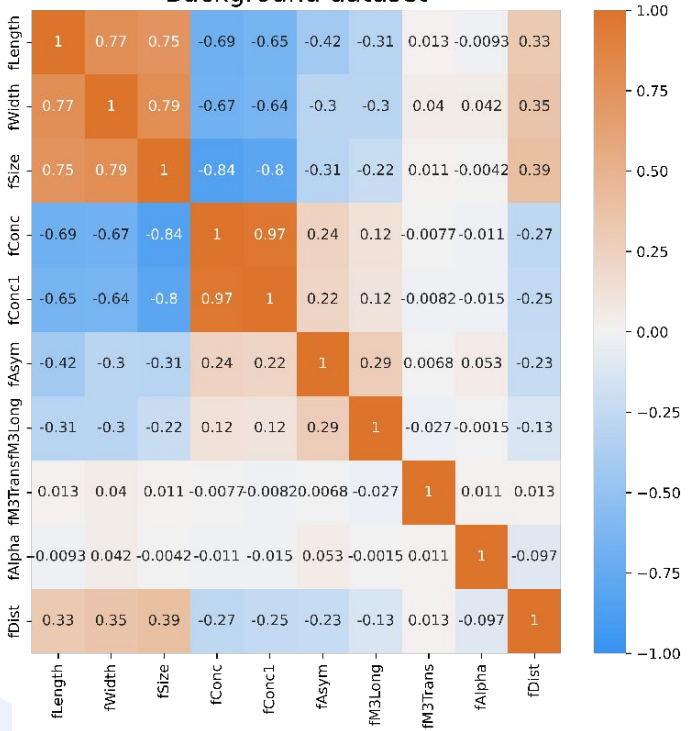
Feature Analysis



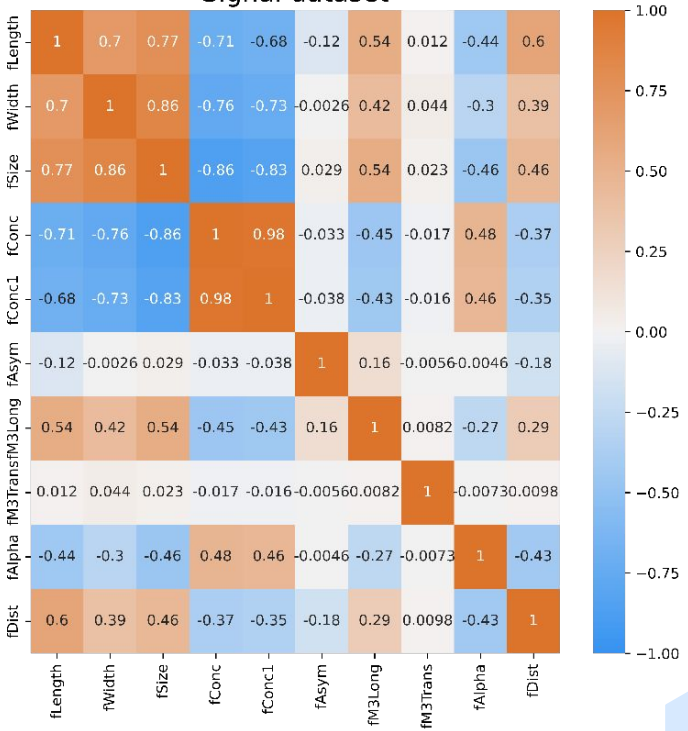
Features Correlation

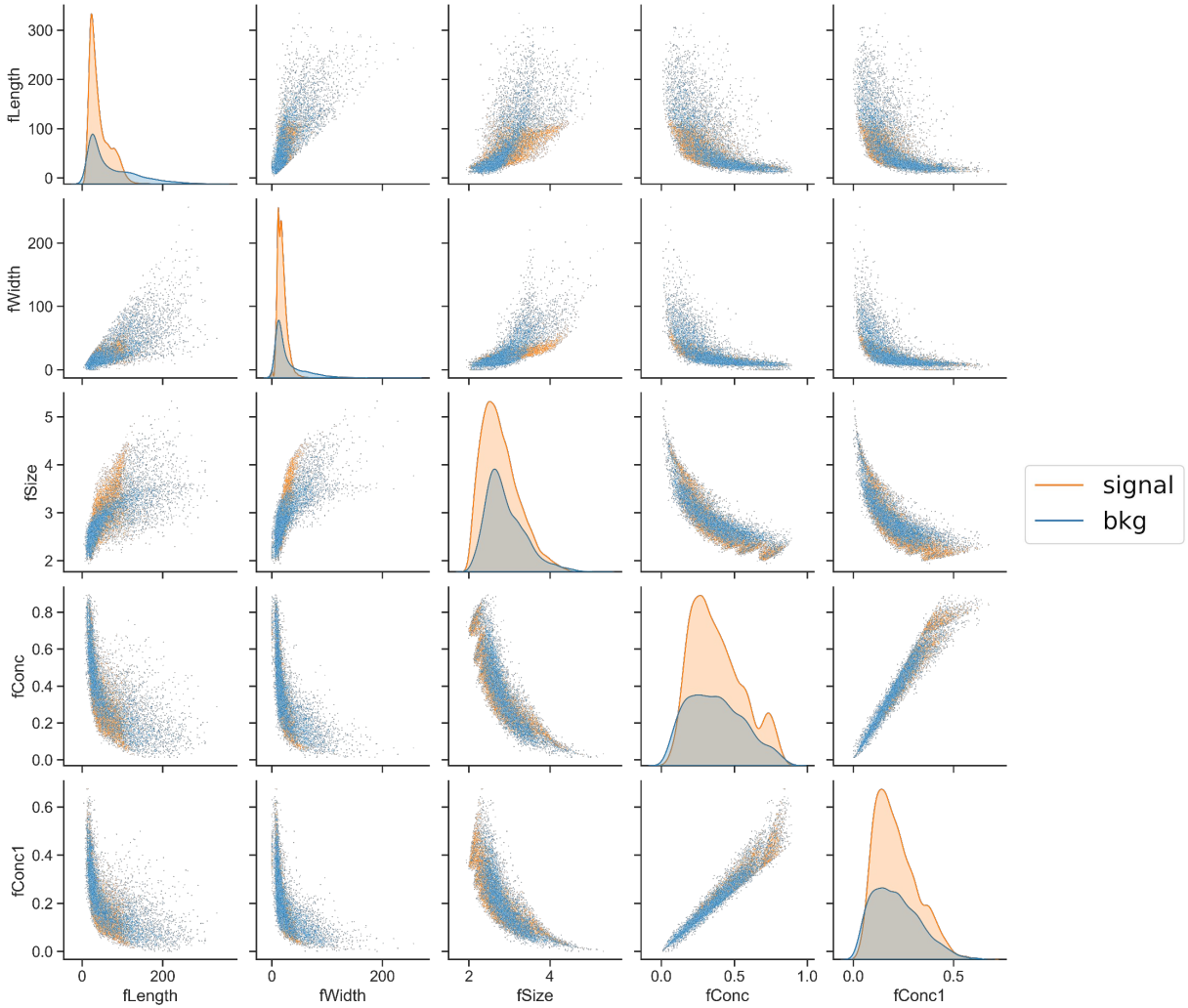
Correlation matrices

Background dataset



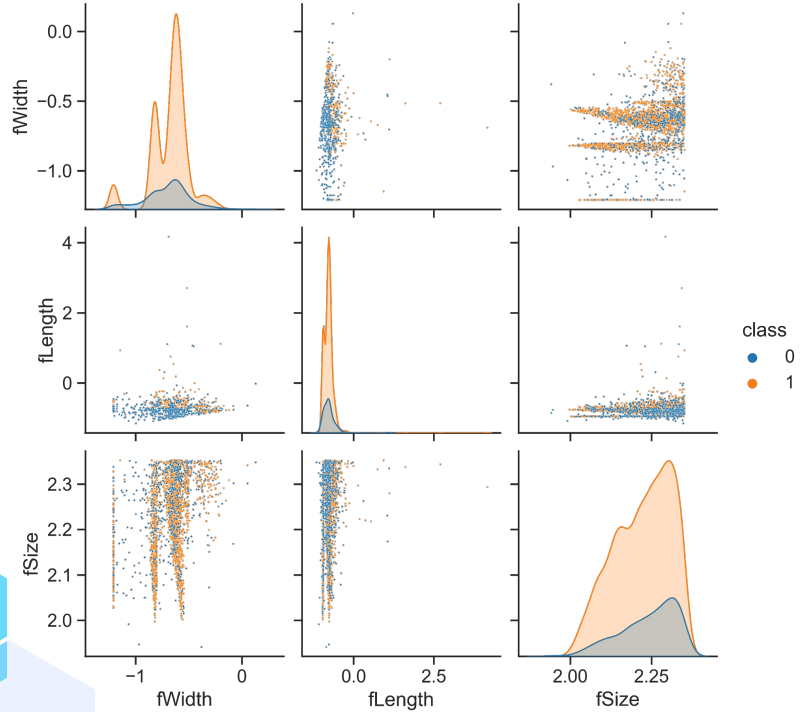
Signal dataset



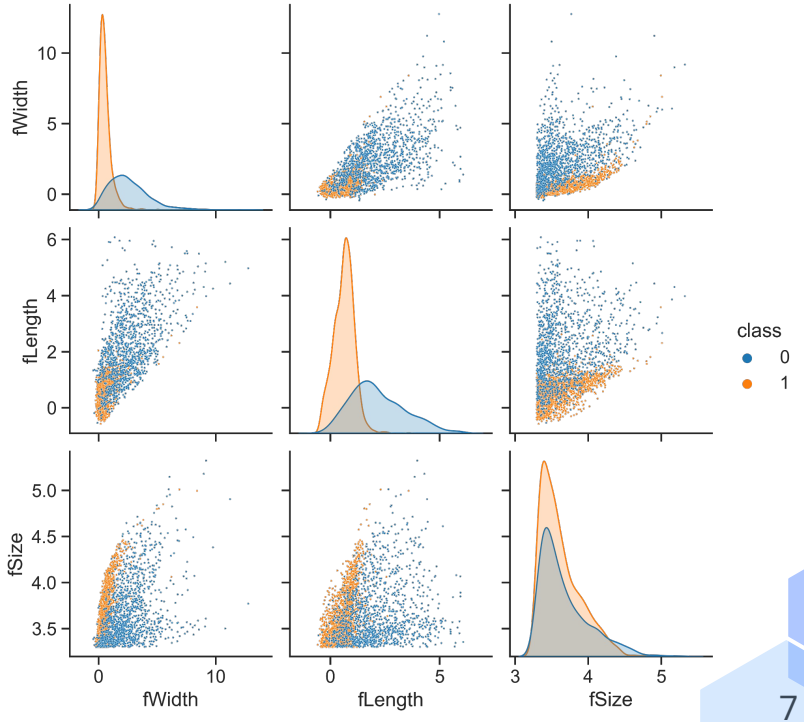

fLength
fWidth
fSize
fConc
fConc1
signal
bkg

Width&Length correlation with energy

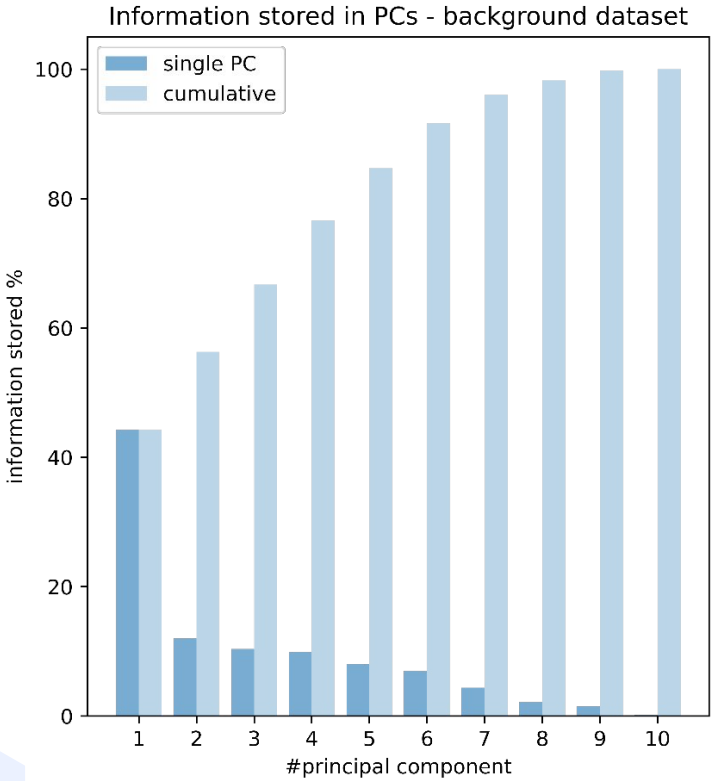
Low photon counting

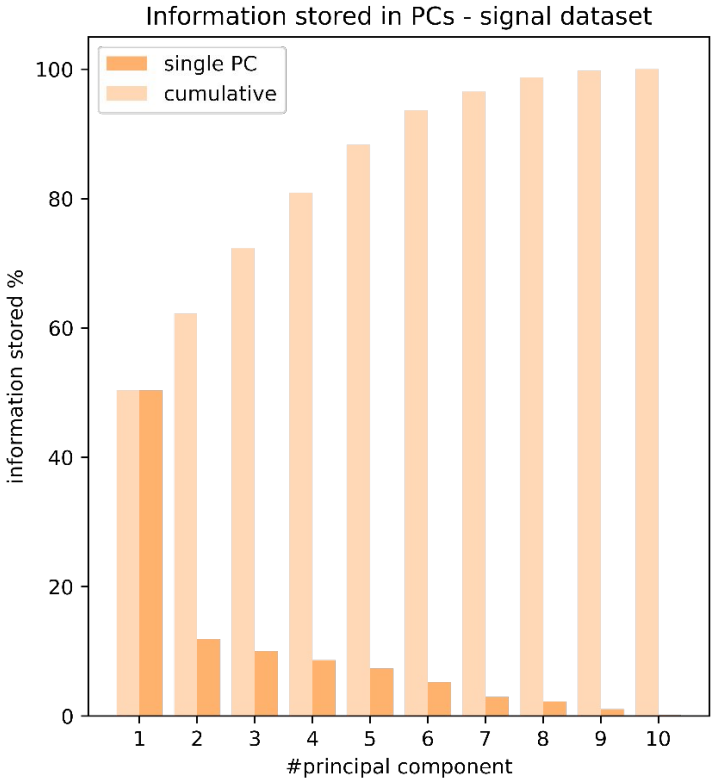


High photon counting



PCA


Information stored in PCs - background dataset
single PC
cumulative
information stored %
0
20
40
60
80
100
1
2
3
4
5
6
7
8
9
10
#principal component


Information stored in PCs - signal dataset
single PC
cumulative
information stored %
0
20
40
60
80
100
1
2
3
4
5
6
7
8
9
10
#principal component

Data Classification

- **Remove** `Alpha` to perform classification task
- **Machine learning** techniques
- **Dataset splitting** and **data standardization**
- **Classifier performance evaluation**

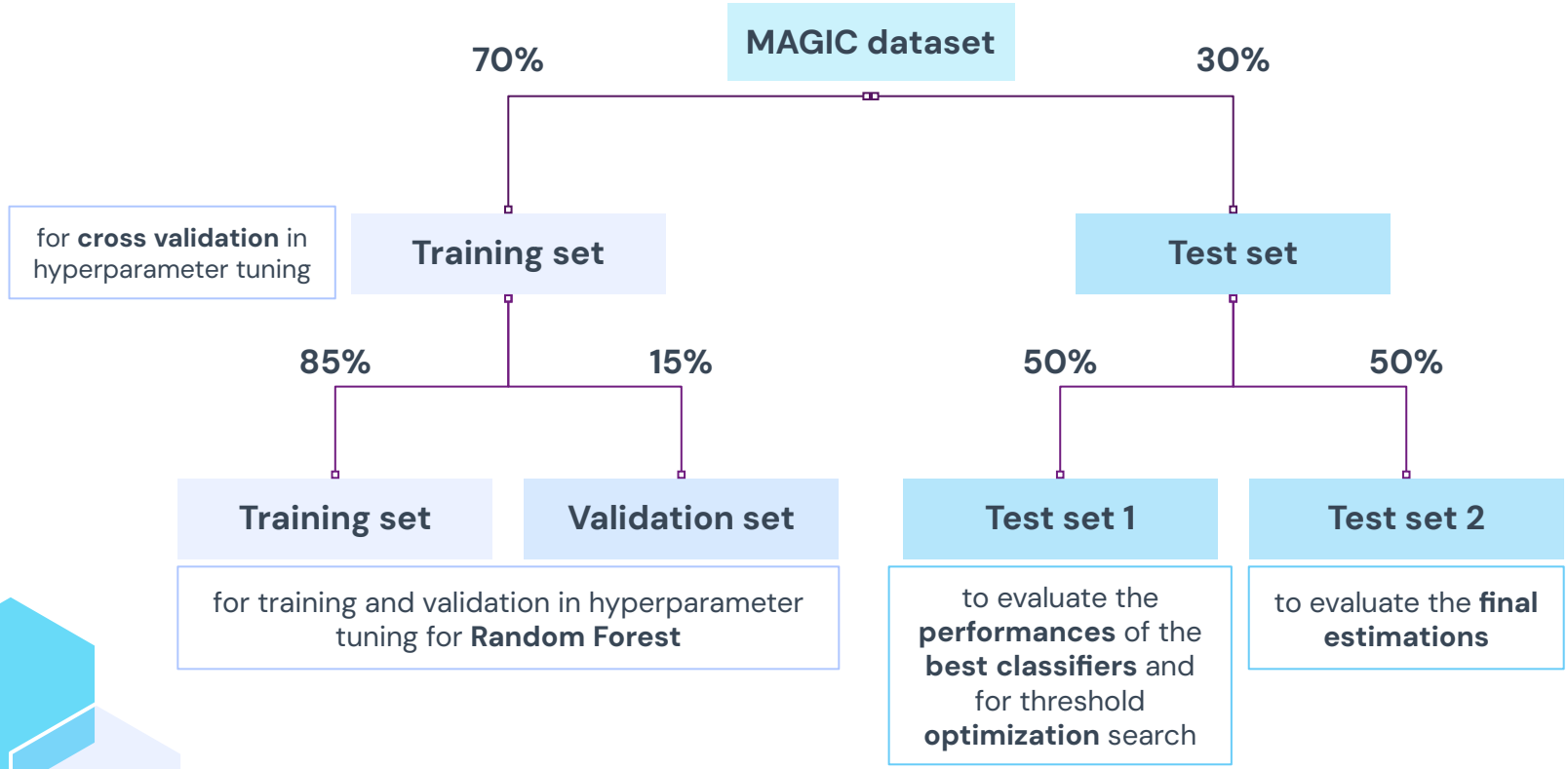
Data Classification

- **Remove** `Alpha` to perform classification task
- **Machine learning** techniques
  - **k-Nearest Neighbors**
  - **SVM** with kernel functions
  - **Neural Network**
  - **Random Forest**
- **Dataset splitting** and **data standardization**
- **Classifier performance evaluation**

Data Classification

- **Remove** `Alpha` to perform classification task
- **Machine learning** techniques
  - **k-Nearest Neighbors**
  - **SVM** with kernel functions
  - **Neural Network**
  - **Random Forest**
- **Dataset splitting** and **data standardization**
- **Classifier performance evaluation**

Dataset splitting



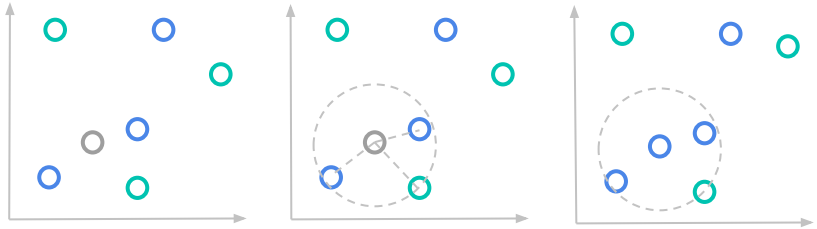
Data Classification

- **Remove** `Alpha` to perform classification task
- **Machine learning** techniques
  - **k-Nearest Neighbors**
  - **SVM** with kernel functions
  - **Neural Network**
  - **Random Forest**
- **Dataset splitting** and **data standardization**
- **Classifier performance evaluation**
  - **Area under ROC curve**
  - **Q - factor** defined as $Q = \frac{\epsilon_\gamma}{\sqrt{\epsilon_h}}$

k-Nearest Neighbors

Grid search: k

Grid search to define the **best number of k-nn**



Other params

Euclidean Metric

Best K-nn model

★ **K-nearest neighbors:** 35

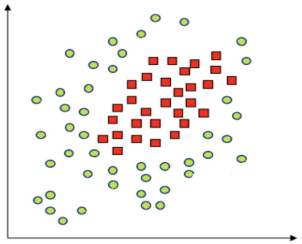
SVM with **kernel methods**

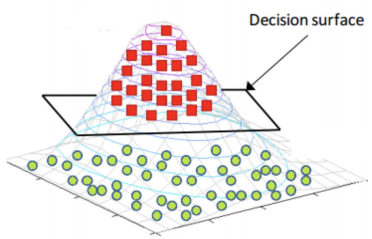
Grid search: Kernel

Gaussian $K(\boldsymbol{x}, \boldsymbol{x}') = e^{-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{2\sigma^2}}$

Polinomial $K(\boldsymbol{x}, \boldsymbol{x}') = (1 + \langle \boldsymbol{x}, \boldsymbol{x}' \rangle)^K$

Grid search to define the **best kernel**





Other params

- ➔ Degree **k** of polynomial
- ➔ **γ** = 1/**σ**

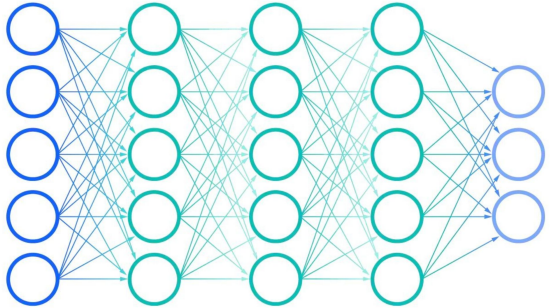
Best kernel model

- ★ **Gaussian** kernel
- ★ **γ:** 0.25

Neural Network

Grid search: Architecture

Grid search to define number of **neurons, layers** and **initial learning rate**



Other fixed params

- ➔ Activation function: **ReLU**
- ➔ **Adaptive** learning rate
- ➔ **L2** regularization
- ➔ Loss: **binary cross entropy**

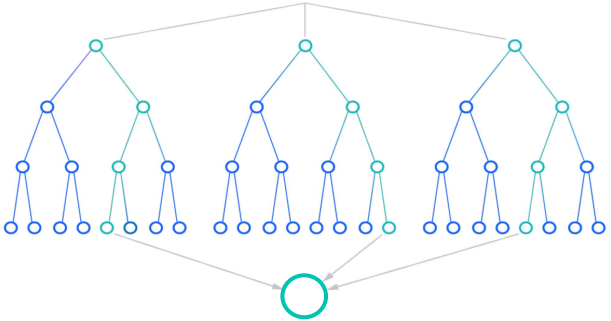
Best NN model

- ★ **Architecture:** (32, 16)
- ★ **Initial Learning Rate:** 0.002

Random Forest

Grid search: Architecture

Grid search to define **number of trees** and **depth**



Other fixed params

- ➔ Features used: **4 random**
- ➔ **Pruning** regularization
- ➔ Loss **Gini index**

Best RF model

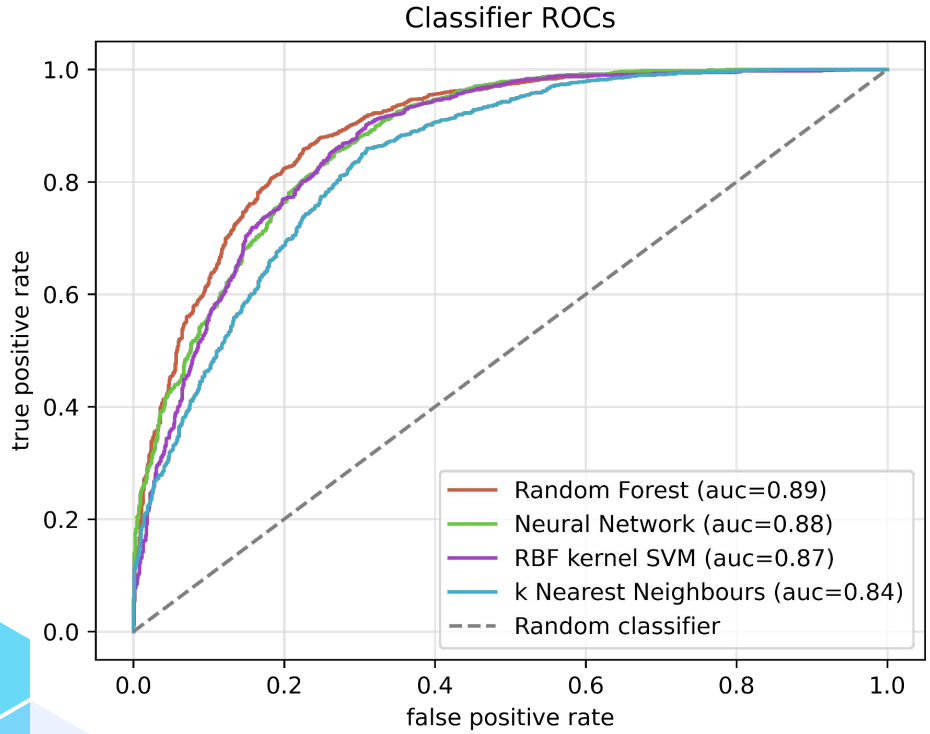
- ★ **Trees:** 70
- ★ **Depth** 20

Classifiers comparison

Summary of the results obtained so far

	K-nearest neighbors	SVM + Kernel	Neural Network	Random Forest
Q factor	1.30	1.46	1.56	**1.59**
Area under ROC	0.84	0.87	0.88	**0.89**

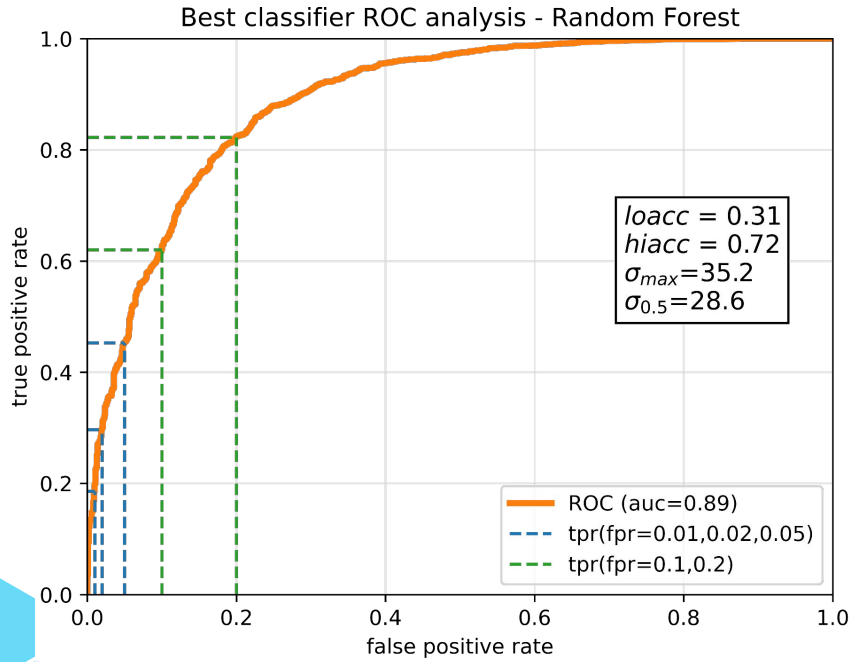
Classifiers comparison



- **Similar results** for the value of the area under the ROC curve
- **Random Forest** classifier returns the highest value and has the best ROC shape

Perform the rest of the analysis using Random Forest as classificator

Random Forest ROC analysis



More metrics:

- ***loacc***: arithmetic mean of true positive rate when false positive rate is 0.01, 0.02, 0.05
- ***hiacc***: arithmetic mean of true positive rate when false positive rate is 0.1, 0.2
- σ_{max}: maximum value of significance computed as

$$\sigma = \frac{S}{\sqrt{2B+S}}$$

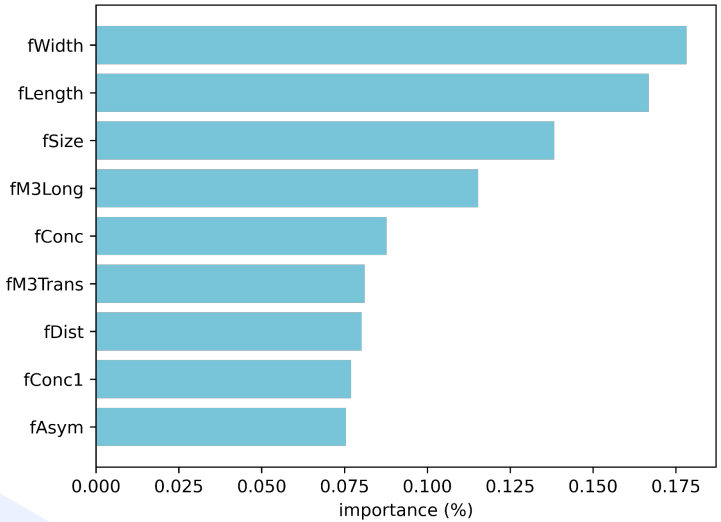
where $S = \epsilon_\gamma N_S$ and $B = \epsilon_h N_B$

- $\sigma_{0.5}$: value of significance found for $\epsilon_\gamma = 0.5$

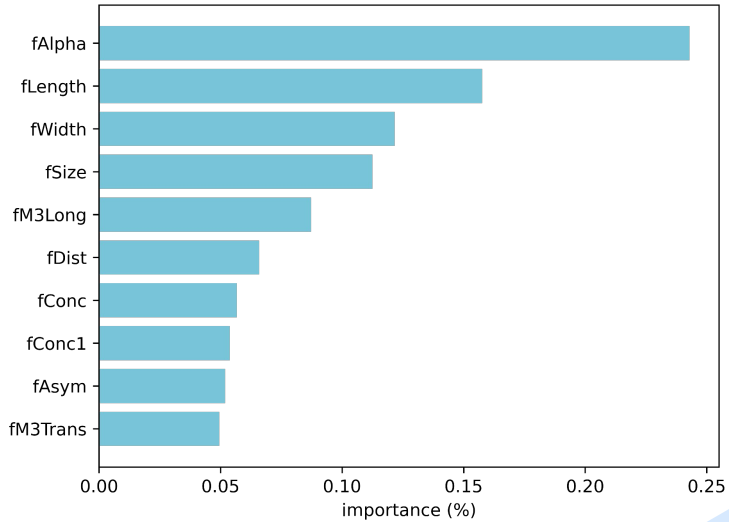
Features importance

Evaluation of the importance of every feature in the Random Forest classifier through **mean decrease impurity** from root to leaves

Without Alpha feature



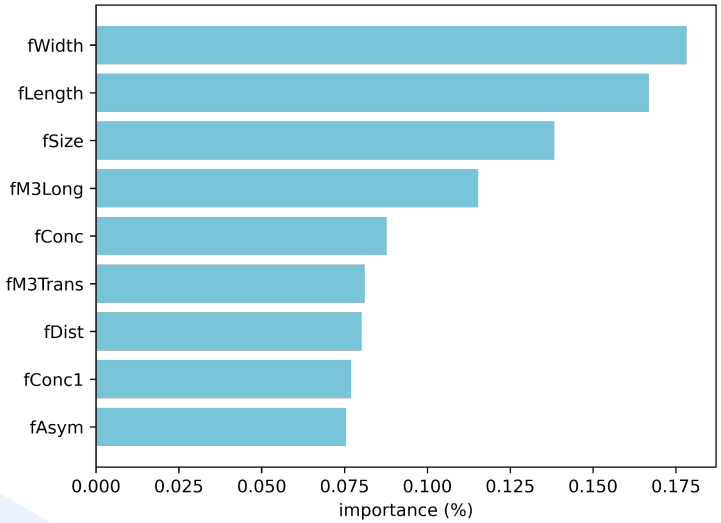
With Alpha feature



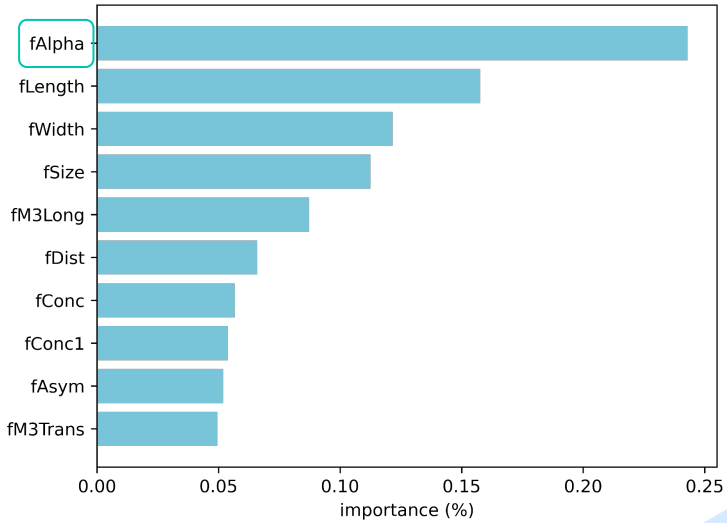
Features importance

Evaluation of the importance of every feature in the Random Forest classifier through **mean decrease impurity** from root to leaves

Without Alpha feature



With Alpha feature



Gammaness & Alpha cuts

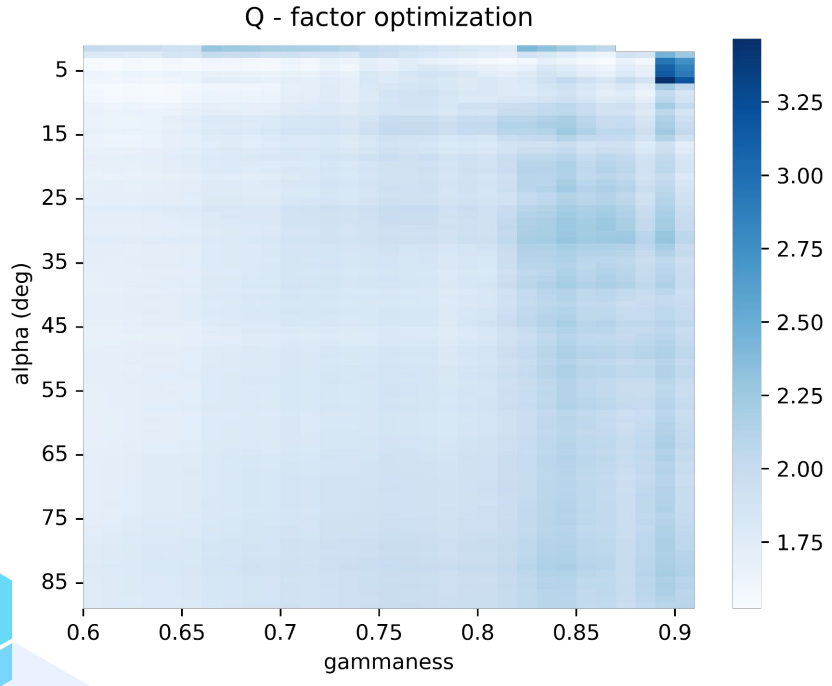
Gammaness
classifier's output, measures the likelihood that an event originated from a primary gamma source

Alpha cut
we can use the `Alpha` variable to set a cut in the data and improve classifier performance since it is the most discriminating feature

we can explore which gammaness threshold AND which alpha cut minimize the false positives rate

Combined search maximizing Q - factor

Gammaness & Alpha cuts



Grid search for Q – factor optimization

- ➔ Gammaness grid step: 0.01
- ➔ Alpha grid step: 1°

Optimal cut values

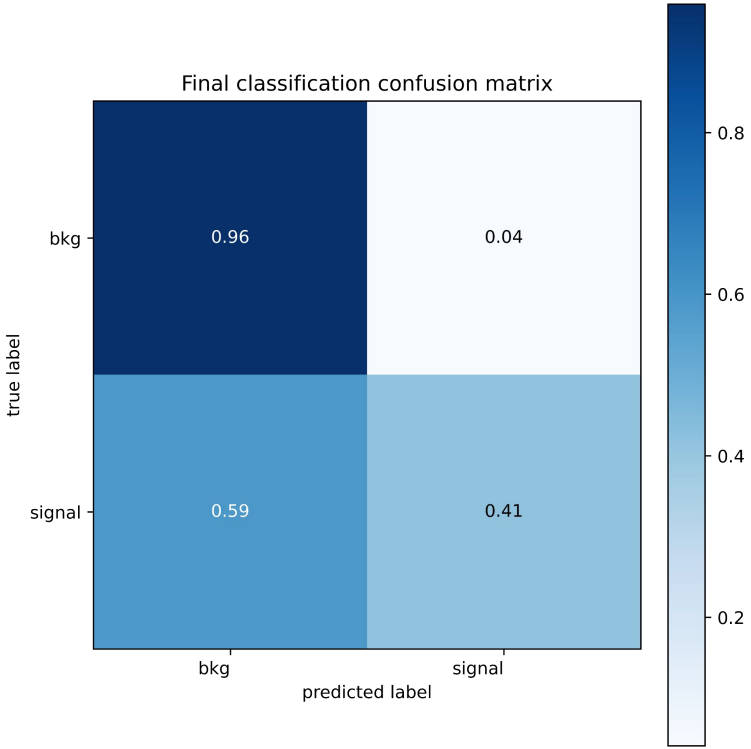
- ➔ Gammaness: 0.89
- ➔ Alpha: 7°

Maximum Q – factor value: 3.5

Gammaness & Alpha cuts

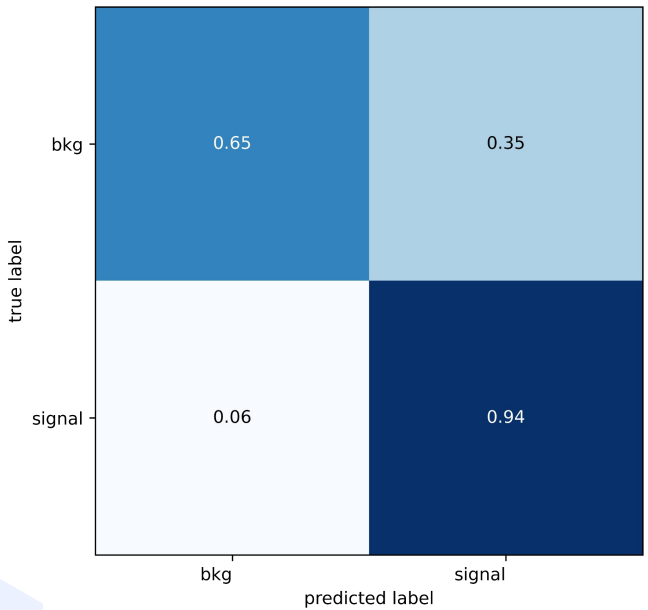
We apply the thresholds to the **test set** to evaluate the final performance:

- ➔ Q-factor: 2.0
- ➔ Area under ROC curve: 0.90
- ➔ σ_{max} : 27
- ➔ $\sigma_{0.5}$: 20

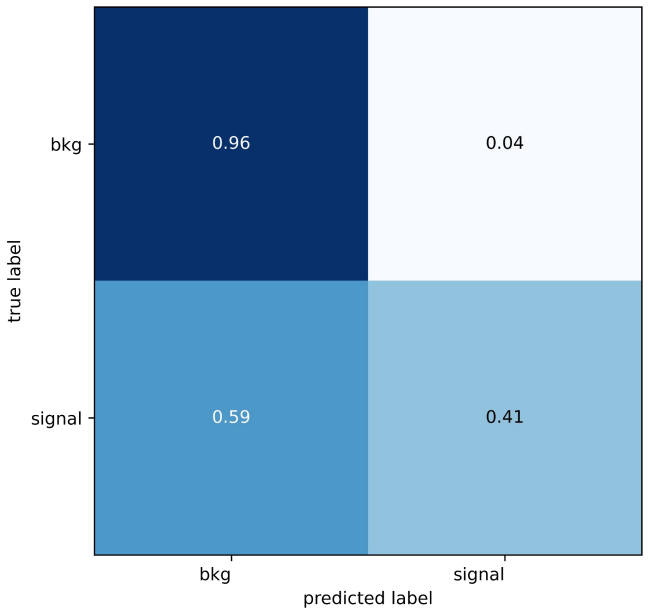


Gammaness & Alpha cuts

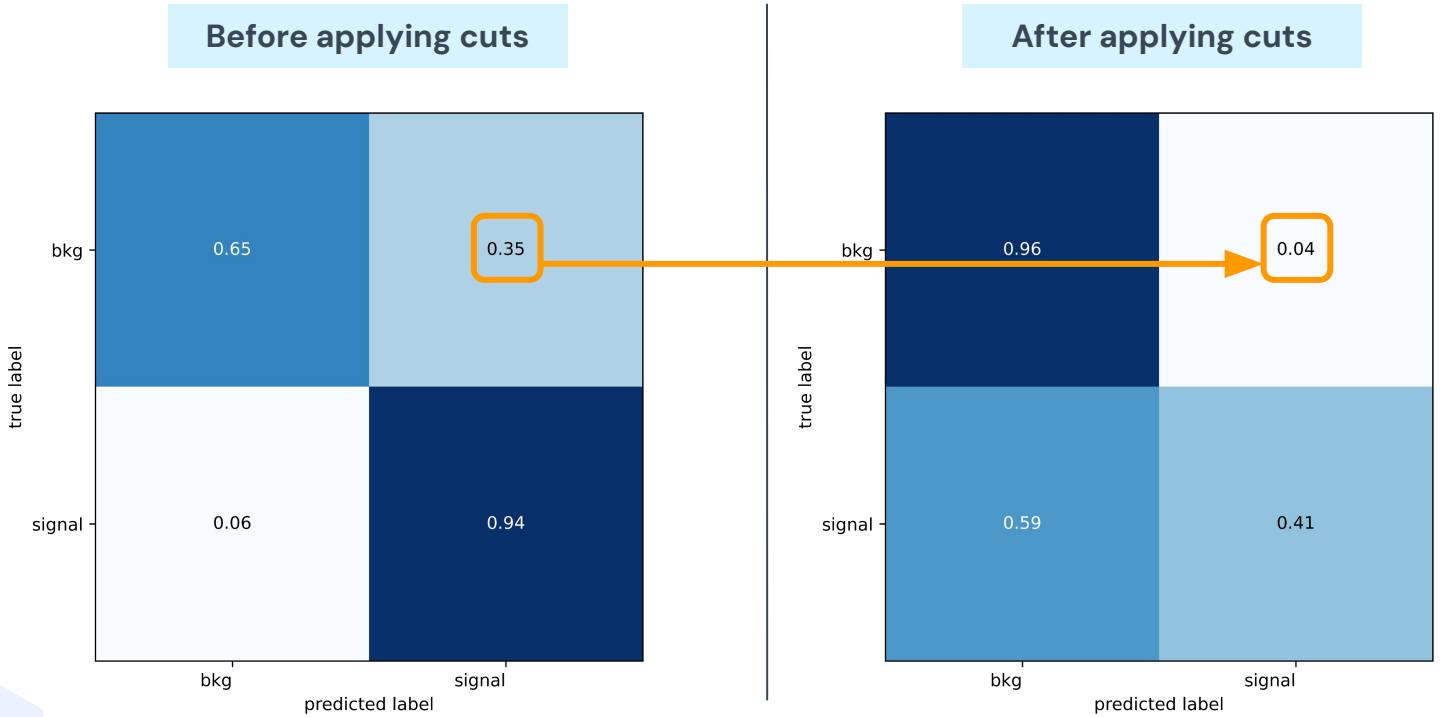
Before applying cuts



After applying cuts



Gammaness & Alpha cuts



Collection Time

The collection time is defined as follows:

$$\Delta t = \frac{n_\gamma}{S \cdot I} \text{ with } I = \int_{0.05TeV}^{50TeV} \frac{dN}{dE} dE$$

with:

$$N = \frac{n_\gamma}{S \cdot \Delta t}$$

$$n_\gamma = TP + FP = 328$$

$$S = 10^9 cm^2$$

➔ Collection time errors are estimated from best fitting parameters' uncertainties.

Collection Time

HEGRA differential spectra models

- Single power law:

$$\frac{dN'_H}{dE} = f'_{0_H} \cdot \left(\frac{E}{E_0}\right)^{-\alpha'_H}$$

with $f'_{0_H} = (2.79 \pm 0.02) \cdot 10^{-11} m^{-2} s^{-1} TeV^{-1}$, $\alpha'_H = 2.59 \pm 0.03$, $E_0 = 1.0 TeV$

- Power law with logarithmic correction on the exponent:

$$\frac{dN''_H}{dE} = f''_{0_H} \cdot \left(\frac{E}{E_0}\right)^{-\alpha''_H + \beta_H log\left(\frac{E}{E_0}\right)}$$

with $f''_{0_H} = (2.67 \pm 0.01) \cdot 10^{-11} m^{-2} s^{-1} TeV^{-1}$, $\alpha''_H = 2.47 \pm 0.10$, $\beta_H = -0.11 \pm 0.1$, $E_0 = 1.0 TeV$

Collection Time

MAGIC differential spectra models

➔ Power law with exponential cut off:

$$\frac{dN'_{MG}}{dE} = f'_{0_{MG}} \cdot \left(\frac{E}{E_0}\right)^{-\alpha'_{MG}} \cdot (e)^{-\frac{E}{E_C}}$$

with $f'_{0_{MG1}} = (3.80 \pm 0.11) \cdot 10^{-11} m^{-2} s^{-1} TeV^{-1}$, $\alpha'_{MG} = 2.21 \pm 0.02$, $E_0 = 1.0 TeV$, $E_C = 6.0 \pm 0.6 TeV$

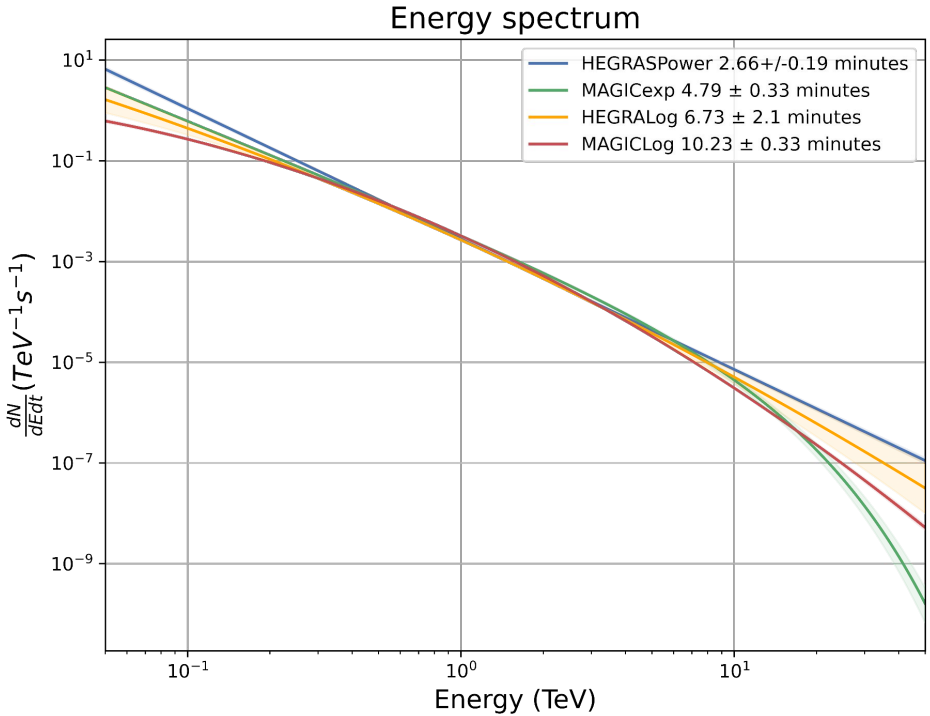
➔ Power law with logarithmic correction on the exponent:

$$\frac{dN''_{MG}}{dE} = f''_{0_{MG}} \cdot \left(\frac{E}{E_0}\right)^{-\alpha''_{MG} + \beta log\left(\frac{E}{E_0}\right)}$$

with $f''_{0_{MG}} = (3.23 \pm 0.03) \cdot 10^{-11} m^{-2} s^{-1} TeV^{-1}$, $\alpha''_{MG} = 2.47 \pm 0.01$, $\beta = -0.24 \pm 0.01$, $E_0 = 1.0 TeV$

Collection Time

Spectra comparison



Conclusions

Dataset characteristics

- ➔ Asymmetric dataset
- ➔ Important features: Alpha, Width, Length, Size, M3Long

Classification results

- ➔ ROC AUC maximizing classifier: Random forests
- ➔ Gammaness/alpha cuts search maximizing Q

Collection time results

- ➔ Best fitting model: MAGIC log-parabola

Thank you for your attention!