Muon Identification Using the XGBoost Toolkit

Maciej Blaut

AGH University of Krakow

Joint ePIC/EICUG Meeting

University of Glasgow, July 11-17, 2026

Contact: `mblaut@student.agh.edu.pl`

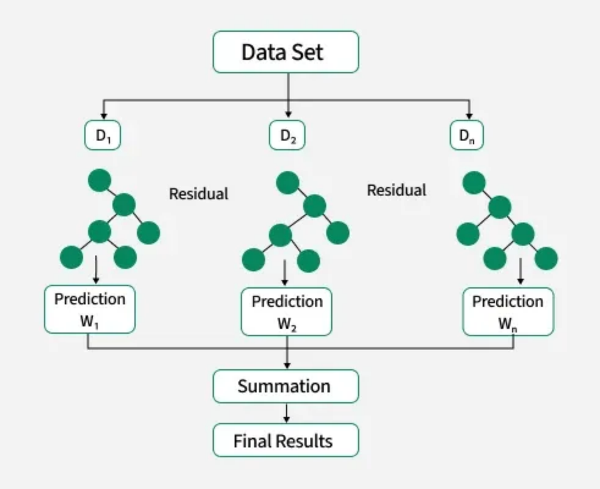
GitHub: github.com/MuonID

XGBoost - eXtreme Gradient Boosting

Core concepts of XGBoost toolkit:

- **Teamwork:** Instead of one complex model, we build hundreds of simple "decision trees" that work together.
- **Learning from mistakes:** Each new tree is designed to fix the errors of the previous ones.
- **Focus on hard cases:** The model doesn't waste time on easy data; it hunts for patterns in the examples it got wrong.
- **Final vote:** The ultimate prediction is a weighted "consensus" from the entire team of trees.

XGBoost working principle:



Source: GeeksforGeeks.org

Data Samples and Methodology

Training:

Single particles from `npsim` particle gun with uniform energy and θ distribution:

- Signal: $\mu^{\pm}$
- Background: $\pi^{\pm}$

Testing:

- **Monte Carlo: GRAPE-Dilepton** Simulated processes: $ep \to ep(\gamma\gamma) \to ep\,\ell^+\ell^-(\gamma)$
- Campaign 26.03: $J/\psi \to \mu^+\mu^-$
- Single: μ, π, e, K, p

Kinematic Cuts applied to ensure high-quality reconstruction:

- **Momentum:** $p > 1$ GeV
- **Pseudorapidity:**
  $|\eta| < 1 \ \vee \ |\eta| > 1.3$ (dead zones/overlap)
  $\eta > -1.25$ (issue with HCal)

These cuts are included in the efficiency calculation.

⋆ *The use of distinct topologies for training and testing ensures robustness and prevents overfitting to specific event geometries.*

Feature Engineering: Raw Calorimetric Features

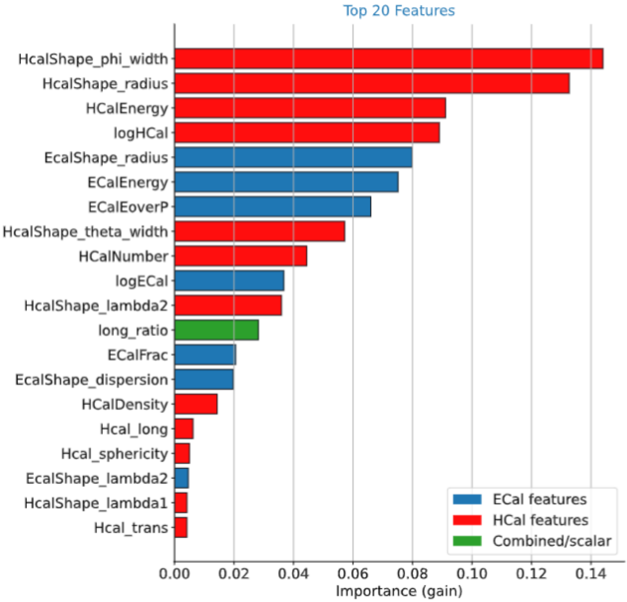
Input for XGBoost Classifier:

- **Global Cluster Statistics:**
  - Total number of clusters per particle.
  - Cumulative energy deposited by a particle in the electromagnetic and hadronic calorimeters.
- **Shower Shape Features** extracted from clusters in both detectors. If more than one cluster was present in a given calorimeter, the clusters were merged.
  Basic geometric descriptors used:
  - **Radius:** Radial extent of the cluster.
  - **Dispersion:** Energy-weighted RMS of hit distances from the cluster center.
  - **2D widths:** Cluster spread in θ and ϕ planes.
  - **3D widths:** Cluster spread in x, y, z coordinates.

Feature Engineering: Data Preprocessing

Feature Derivation & Utility

- **Preprocessing:** From those 20 features, derived features are created via mathematical transformations.
- **Purpose:** These non-linear transformations better expose selection patterns and enhance classifier sensitivity.
- **Reference:** Complete feature specifications are provided in the backup slides.



Currently used method

Cuts were applied to the ratios of total energy deposited in the calorimeter to the momentum:

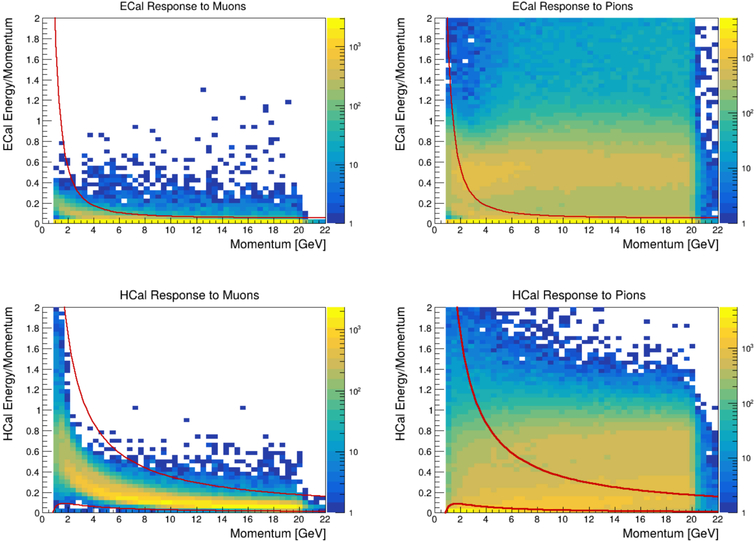
$$\frac{\sum_i E_{ECal}}{p}, \qquad \frac{\sum_i E_{HCal}}{p}.$$

Cuts on PID information from different detectors were not used

Performance of the method (uniform distribution):

- Muon Efficiency: 97.7%
- Pion Rejection: 86.75%

E/p ratios as a function of momentum:



XGBoost Response Selection

Classifier Working Point:

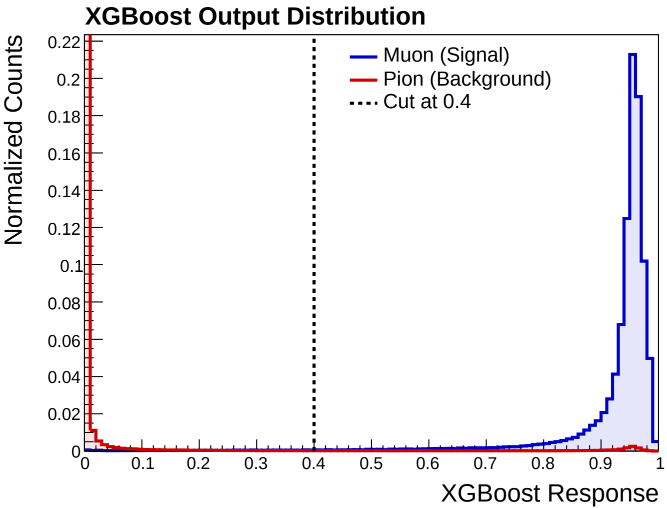
- A final selection cut was applied to the XGBoost response to maximize the F_1-score.
- **Selected Threshold:** `Response > 0.40`

Performance of the method (uniform distribution):

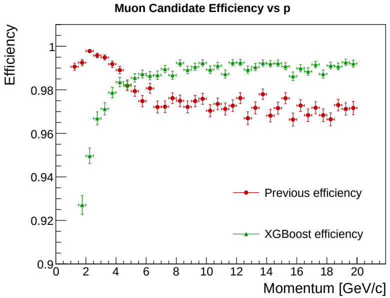
- **Muon Efficiency:** 98.0%
- **Pion Rejection :** 96.3%

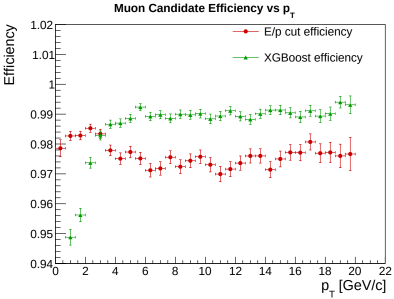
F_1-score

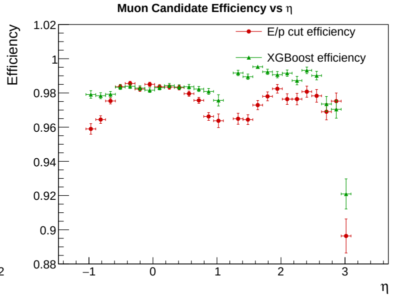
Harmonic mean of the precision and recall.

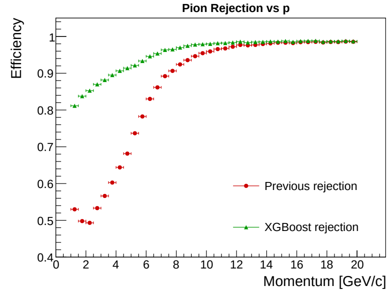


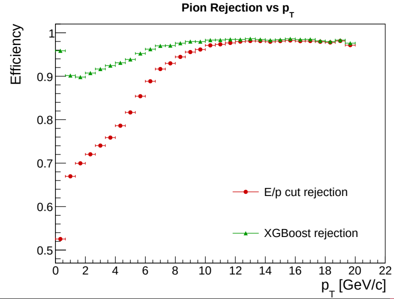
Muon Identification Efficiency

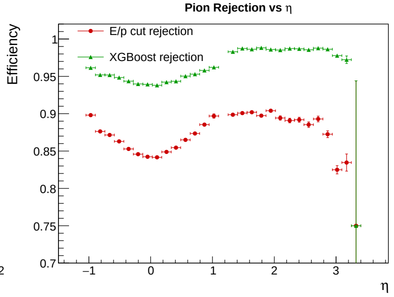












Muon Identification Performance: Current method vs XGBoost

The final total muon identification efficiency and background rejection, including information from RICH:

	Single particle		J/ψ sample / pion sample	
Method	**Muon Eff.**	**Pion Rej.**	**Muon Eff.**	**Pion Rej.**
Current	97.7%	98.7%	95.3%	98.8%
XGBoost	**98.3%**	**99.2%**	**91.7%**	**99.8%**

Key Findings:

- **Observed Limitations:** Lower efficiency for **low**-pT **muons**.
- **Advantage:** Higher pion rejection in **low**-p,which is crucial for J/ψ analysis.
- **XGBoost Advantage:** There is a consistent improvement in both muon identification efficiency and background rejection compared to the current selection.
- **Hadron Suppression:** Kaons and protons exhibit rejection rates similar to those observed for pions.

Future Work and Outlook

- **Model Optimization & Refinement:**
  - Search for the optimal hyperparameters,
- **Detector & Physics Studies:**
  - Improving the RICH detector output,
  - Adding the information from ToF and Tracker.

Thank you for your attention!

Backup slides

ECalShape Features

- EcalShape_radius - Cluster radius
- EcalShape_dispersion - Energy dispersion
- EcalShape_theta_width - Theta spread
- EcalShape_phi_width - Phi spread
- EcalShape_lambda1 - First eigenvalue - x
- EcalShape_lambda2 - Second eigenvalue - y
- EcalShape_lambda3 - Third eigenvalue - z

HCalShape Features

- HcalShape_radius - Cluster radius
- HcalShape_dispersion - Energy dispersion
- HcalShape_theta_width - Theta spread
- HcalShape_phi_width - Phi spread
- HcalShape_lambda1 - First eigenvalue - x
- HcalShape_lambda2 - Second eigenvalue - y
- HcalShape_lambda3 - Third eigenvalue - z

Backup slides

Scalars (Energy/Density)

- ECalEnergy/HCalEnergy - Raw energy
- ECalNumber/HCalNumber - Hit counts
- ECalEoverP/HCalEoverP - E/p ratio
- ECalFrac/HCalFrac - E / (E1+E2)
- HitRatio - ECalN / HCalN
- ECalDensity/HCalDensity - E / N
- EoverP_ratio - ECalEp / HCalEp
- logECal/logHCal - log1p(E)

Calculated Shapes

- Ecal_trans/Hcal_trans - $\sqrt{\lambda_1 \lambda_2}$
- Ecal_long/Hcal_long - λ_3
- Ecal_LoverT/Hcal_LoverT - λ_3/trans
- Ecal_sphericity/Hcal_sphericity - λ_1/λ_3
- Ecal_angular_asym/Hcal_angular_asym - $(\theta - \phi)/(\theta + \phi)$
- radius_ratio/disp_ratio - E/H ratio
- trans_ratio/long_ratio - E/H ratio
- LoverT_mismatch - abs(diff)
- sphericity_mismatch - abs(diff)
- Radial_HCal_Fraction - HcalR/(EcalR+HcalR)

Total: 45 features were used in the model.

Parameters Distribution