Measurement of the Photon Identification Efficiency with the Matrix Method Using Neural Networks

Can Süslü

Physikalisches Institut,
University of Bonn, Germany

can.suslu@ug.bilkent.edu.tr

October 7, 2022

Outline

Motivation

Motivation

- Photon Identification efficiency is important for: **inclusive prompt photon** and **di-photon** cross section, measurement of $H \to \gamma\gamma$ or any process that involves prompt photons in the final state.

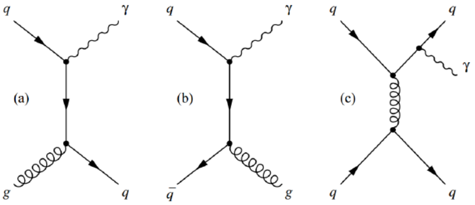


Figure 1: Prompt photon production via a)quark-gluon scattering b)quark-anti-quark annihilation c)bremsstrahlung radiation off of an outgoing quark.

- Fake photons consists of photons originated from the jets from the neutral hadron decays (π^0, η mesons), and misreconstructed e^-e^+ pairs.

- **Matrix Method** is a data-driven method to measure the photon ID efficiency.
    - **Narrow-strip** and **relaxed-tight variables** as DVs (track isolation variables).

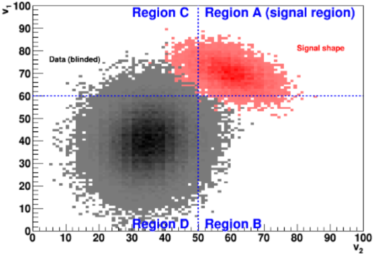


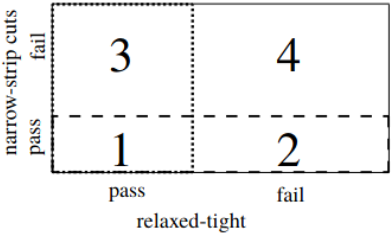
Figure 2: Canonical example of ABCD Method.

- Ideally, the matrix method is suitable for uncorrelated DVs. However, the photon isolation and shower shape variables are correlated.
- Instead of using rectangular cuts, **two neural networks** are used to determine the pass/fail statement for each narrow-strip and relaxed tight variables.
- Neural network outputs are used for the phase space separation in the matrix method.

Matrix Method

Matrix Method

- Matrix method uses narrow-strip, and relaxed tight variables.



- $\epsilon^{tight-ID} \equiv \frac{N^s_{ID}}{N^s}$
- $\hat{N}_a$: Number of track isolated photons in region a.
- $N^T_{ID} = N^b_{ID} + N^s_{ID}$

$$\epsilon^{tight-ID} = \frac{\frac{\hat{\epsilon_{ID}} - \hat{\epsilon^b_{ID}}}{\hat{\epsilon^s_{ID}} - \hat{\epsilon^b_{ID}}} \cdot N^T_{ID}}{\frac{\hat{\epsilon} - \hat{\epsilon^b}}{\hat{\epsilon^s} - \hat{\epsilon^b}} \cdot N^T} \qquad (1)$$

$$\hat{\varepsilon}^b_{ID} = \frac{\hat{N}^b_1}{N^b_1} \approx \frac{\hat{N}^b_3}{N^b_3} = \frac{R_p \cdot \hat{\varepsilon}_3 - A \cdot f_p \cdot \hat{\varepsilon}^s_3}{R_p - A \cdot f_p} \qquad (2)$$

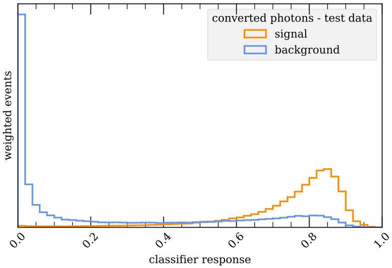
$$\hat{\varepsilon}^b = \frac{\hat{N}^b_{1+2+3+4}}{N^b_{1+2+3+4}} \approx \frac{\hat{N}^b_{2+3+4}}{N^b_{2+3+4}} = \frac{R_a \cdot \hat{\varepsilon}_{2+3+4} - A \cdot f_a \cdot \hat{\varepsilon}^s_{2+3+4}}{R_a - A \cdot f_a} \qquad (3)$$

- This assumption leads to systematic uncertainties.(Correlation between narrow strip and track isolation.)

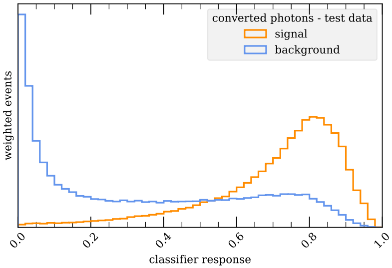
Analysis

- In the analysis, the ATLAS data from 2015 to 2018, signal and background MC data are used.
- The data are put into the neural networks which outputs the pass/fail rate between 0-1.

Narrow Strip Variables



Relaxed Tight Variables



NN Parameters

Narrow Strip NN

- Training Variables: w_{η_1}, ΔE_s, f_{side}, E_{ratio}, f_1, e_{277}, p_T, η

ECAL first layer

Relaxed Tight NN

- Training Variables: R_η, R_ϕ, w_{η_2}, $s_{,tot}$, R_{had}, R_{had_1}, p_T, η

ECAL second layer + Hadronic Leakage

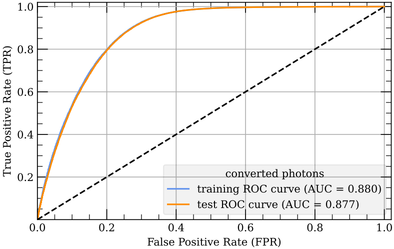


Figure 3: ROC Curve of Narrow Strip NN

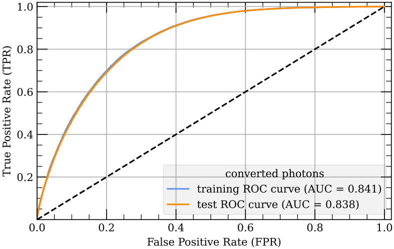


Figure 4: ROC Curve of Relax Tight NN

The analysis script classifies the events into 4-different regions according to the $y_{prediction}$.

- Converted photons
- Track Isolation Criteria:

$$\text{topoEtcone20} < 6.5 \cdot 10^{-2} \cdot E_T \text{ and } \mathrm{ptCone}2\theta < 0.05 \cdot E_T$$

- 4 $|\eta|$ intervals:

$$[0.00, 0.60), [0.60, 1.37), (1.52, 1.81), [1.81, 2.37)$$

- 13 unequal E_T bins from 25 to 1500 GeV.
- Pass threshold for Relax Tight: $y_{pred} > 0.55$
- Pass threshold for Narrow Strip: $y_{pred} > 0.6$

Calculate the following quantities,

- from the ATLAS data: R_p,R_a,$\hat{\epsilon}_{ID}$, $\hat{\epsilon}$, $\hat{\epsilon}_3$, $\hat{\epsilon}_{2+3+4}$
- from the Signal MC: f_p,f_a,$\hat{\epsilon}^s_{ID}$, $\hat{\epsilon}^s$, $\hat{\epsilon}^s_3$, $\hat{\epsilon}^s_{2+3+4}$

and save them into **json** files.
The script saves the root files into a Pandas dataframe, and creates dictionaries for each region. The data is stored for each η-p_T bin.

Systematic Uncertainties

- Sources of systematic uncertainties: MC statistics, track isolation requirement, detector material, closure uncertainty.

Closure Test

$$\Delta\hat{\varepsilon}^b_{ID} = \frac{\left|\hat{\varepsilon}^b_1 - \hat{\varepsilon}^b_3\right|}{\hat{\varepsilon}^b_1} \tag{4}$$

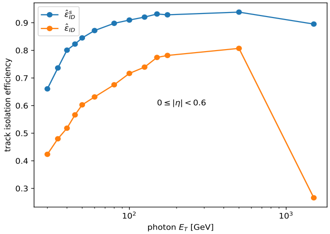
$$\Delta\hat{\varepsilon}^b = \frac{\left|\hat{\varepsilon}^b_{1+2+3+4} - \hat{\varepsilon}^b_{2+3+4}\right|}{\hat{\varepsilon}^b_{1+2+3+4}} \tag{5}$$

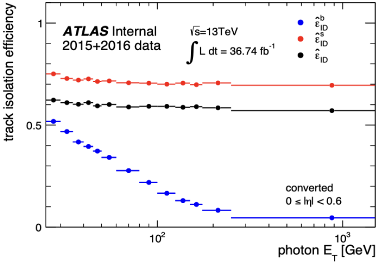
	converted photons	
$\|\eta\|$ interval	$\Delta\hat{\varepsilon}^b_{ID}[\%]$	$\Delta\hat{\varepsilon}^b[\%]$
$[0.0, 0.6)$	7.55	10.01
$[0.6, 1.37)$	10.24	11.36
$(1.52, 1.81)$	5.45	11.40
$[1.81, 2.37)$	3.36	13.73

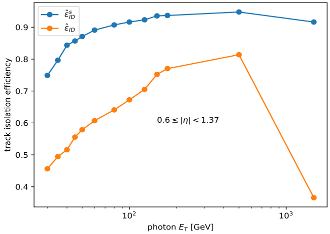
(6)

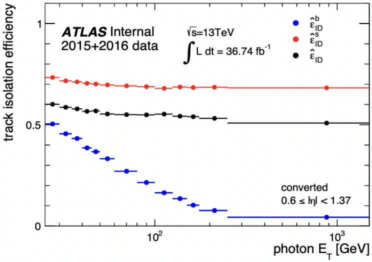
Results

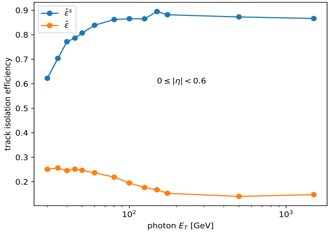
Results for Track Isolation Efficiencies

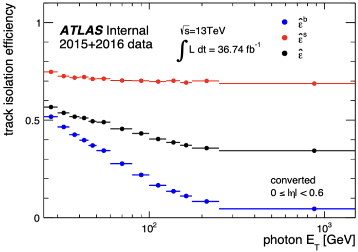


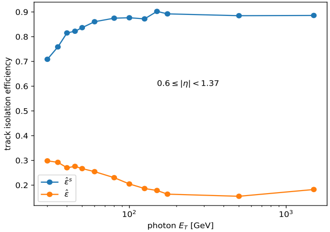


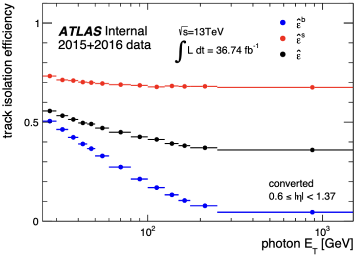





$\hat{\varepsilon}^s$
$\hat{\varepsilon}$
$0 \leq |\eta| < 0.6$
track isolation efficiency
photon E_T [GeV]


ATLAS Internal
2015+2016 data
√s=13TeV
∫L dt = 36.74 fb⁻¹
$\hat{\varepsilon}^b$
$\hat{\varepsilon}^s$
$\hat{\varepsilon}$
converted
$0 \leq |\eta| < 0.6$
track isolation efficiency
photon E_T [GeV]

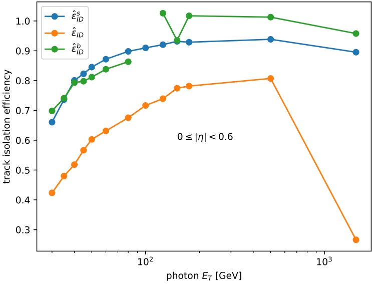

$0.6 \leq |\eta| < 1.37$
$\hat{\varepsilon}^s$
$\hat{\varepsilon}$
track isolation efficiency
photon E_T [GeV]

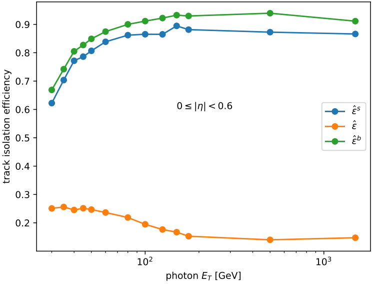

ATLAS Internal
2015+2016 data
√s=13TeV
∫L dt = 36.74 fb⁻¹
$\hat{\varepsilon}^b$
$\hat{\varepsilon}^s$
$\hat{\varepsilon}$
converted
$0.6 \leq |\eta| < 1.37$
track isolation efficiency
photon E_T [GeV]

To obtain the track isolation efficiency for the background sample:

$$a \cdot \left(\hat{\varepsilon}^b\right)^2 - b \cdot \hat{\varepsilon}^b + c = 0 \tag{7}$$

$$\begin{aligned} a &= R_a - f_a \\ b &= R_a \cdot \left(\hat{\varepsilon}^S + \hat{\varepsilon}_{2+3+4}\right) - f_a \cdot \left(\hat{\varepsilon} + \hat{\varepsilon}^S_{2+3+4}\right) \\ c &= R_a \cdot \hat{\varepsilon}_{2+3+4} \cdot \hat{\varepsilon}^s - f_a \cdot \hat{\varepsilon}^S_{2+3+4} \cdot \hat{\varepsilon} \end{aligned} \tag{8}$$

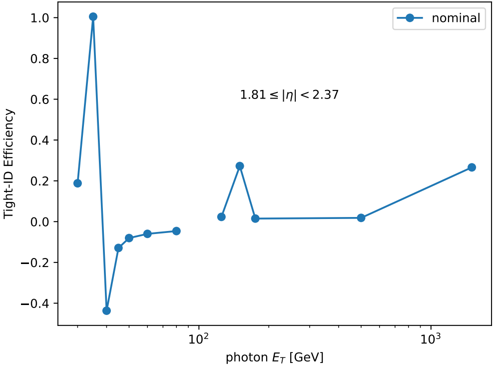




Tight Identification Efficiency

$$\varepsilon^{tight-ID}\left(N^{T}_{ID}, N^{T}, \hat{\varepsilon}^{S}_{ID}, \hat{\varepsilon}^{S}, \hat{\varepsilon}^{b}_{ID}, \hat{\varepsilon}^{b}, \hat{\varepsilon}_{ID}, \hat{\varepsilon}\right) = \frac{\frac{\hat{\varepsilon}_{ID}-\hat{\varepsilon}^{b}_{ID}}{\hat{\varepsilon}^{s}_{ID}-\hat{\varepsilon}^{b}_{ID}} \cdot N^{T}_{ID}}{\frac{\hat{\varepsilon}-\hat{\varepsilon}^{b}}{\hat{\varepsilon}^{s}-\hat{\varepsilon}^{b}} \cdot N^{T}} \tag{9}$$

- High fluctuation due to **systematic errors**, and maybe due to a bug inside the code.



Conclusion

Conclusion

- The outputs of NNs are used instead of a cut based method.
- Matrix Method is an easy way to measure the photon ID efficiency.
- A further NN optimization should be done in the future for a better seperation, and for lower systematic uncertainties.
- A script that uses Uproot, Pandas, and numpy is written. Tensorflow is used for the neural network.
- Hands-on experience on working with real data, using modern ML techniques, a great motivation to start Master studies !...

That's it from my side, thanks for listening!

Backup Slides

Category	Variable name	Description
Hadronic leakage	R_{had_1}	Ratio of E_T in the first sampling layer of the hadronic calorimeter to E_T of the EM cluster (used over the range $\lvert\eta\rvert < 0.8$ or $\lvert\eta\rvert > 1.52$)
	R_{had}	Ratio of E_T in the hadronic calorimeter to E_T of the EM cluster (used over the range $0.8 < \lvert\eta\rvert < 1.37$)
ECAL first layer	w_{η_1}	Lateral shower width, $\sqrt{\sum E_i(i-i_{\text{max}})^2/\sum E_i}$, where i runs over all strips in a window of 3×2 $\eta\times\phi$ strips, and i_{max} is the index of the highest-energy strip calculated from three strips around the strip with maximum energy deposit
	$w_{s,\text{tot}}$	Total lateral shower width $\sqrt{\sum E_i(i-i_{\text{max}})^2/\sum E_i}$, where i runs over all strips in a window of 20×2 $\eta\times\phi$ strips, and i_{max} is the index of the highest-energy strip measured in the strip layer
	f_{side}	Energy outside the core of the three central strips but within seven strips divided by energy within the three central strips
	ΔE_s	Difference between the energy associated with the second maximum in the strip layer and the energy reconstructed in the strip with the minimum value found between the first and second maximum
	E_{ratio}	Ratio of the energy difference between the maximum energy deposit and the energy deposit in the secondary maximum in the cluster to the sum of these energies
	f_1	Ratio of the energy in the first layer to the to the total energy of the EM cluster
ECAL second layer	R_η	Ratio of the energy in 3×7 $\eta\times\phi$ cells over the energy in 7×7 cells centered around the photon cluster position
	w_{η_2}	Lateral shower width, $\sqrt{\sum E_i\eta_i^2/\sum E_i - (\sum E_i\eta_i/\sum E_i)^2}$, where E_i is the energy and η_i is the pseudorapidity of cell i and the sum is calculated within a window of 3×5 cells
	R_ϕ	Ratio of the energy in 3×7 $\eta\times\phi$ cells over the energy in 3×7 cells centered around the photon cluster position

Classifier Settings/Hyperparameters:

- 4 Layers with size of 32.
- Batch Size: 8192
- Epochs: 250
- Learning Rate: 0.001
- Adam Optimizer