EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

CERN-EP-XXXX-XXX
Day Month XXXX

Identification of Prompt Photons with the ALICE detector

Anders Sandermann Mortensen

Abstract

In this project, two different methods for identifying prompt photons have been tested. The baseline method was a classical cuts-based analysis. This set a benchmark performance for novel methods. One new method was tested here, employing a boosted decision tree to perform the identification. The machine learning approach showed a significant improvement in purity and efficiency compared to the benchmark. With further development this can be a very robust and efficient method for identifying prompt photons in future analysis.

1 Identification of prompt photons

Prompt photons are photons which are created in the initial, hard scattering stage of a particle collision. These kinds of interactions are signified by high momentum transfer. To leading order, prompt photons are mainly produced in Compton scattering in $gq \rightarrow \gamma g$ processes. Including next to leading order terms, they can also be produced in fragmentation processes. In those cases, the initial hard scattering process will produce, for example, $gq \rightarrow \gamma gg$ or $qq \rightarrow \gamma g$. What makes these photons so attractive to study is the fact that their dominant source is quark-gluon interactions. Therefore, they are sensitive to the gluon parton distribution function, which is not well constrained. In addition, photons do not interact strongly, so should a strongly coupled QGP (or a similar phase of matter) be created in the collision, the photons will be largely unaffected by it[1].

2 Experimental setup

The prompt photons are studied here with data from ALICE at CERN. ALICE is a dedicated heavy-ion collision experiment consisting of multiple subdetectors. The detectors relevant to this analysis are The electromagnetic calorimeter (EMCal) and the tracking detectors (ITS and TPC). Additional detectors are used for triggering etc. but they are not directly relevant for this analysis. The EMCal is a sandwich-type calorimeter built from alternating layers of lead and scintillator material. When ionising particles hit the EMCal, their energy is absorbed by the lead and scintillator layers. The scintillator layers produce light. Via wave-length shifting fibres, this light moves towards an avalanche photo-diode that can measure it. This allows the particle energy to be measured. The main tracking detectors of ALICE are the ITS and the TPC. The ITS consists of multiple layers of silicon detectors, allowing for exact and fast tracking close to the ALICE interaction point. The TPC is a time projection chamber that can track particles very precisely. Details on the properties of the ALICE detectors can be found in the relevant TDR's[2]

Charged particles will leave tracks in the ALICE tracking detectors and clusters in the EMCal. However, photons will only appear as clusters in the EMCal, but not in the tracking detectors as they are neutral. The tracking detectors can, therefore, aid in determining which clusters are from photons and which are not.

3 Baseline analysis

The baseline cuts-based analysis was previously done in $\sqrt{s_{NN}} = 8$TeV pp and $\sqrt{s_{NN}} = 5.02$TeV $p-Pb$-collision data by F. Jonas in [1]. The same set of cuts were applied again in this analysis. The motivation for repeating the analysis was partly to determine the purity and efficiency of this identification approach for a new collision system energy $\sqrt{s_{NN}} = 13$TeV pp-collisions. Secondly, this process was part of a larger project, where the cuts-based analysis was implemented into a more streamlined analysis-framework.

The efficacy of the various identification schemes will be measured by comparing the purity and efficiency achieved with different methods. As mentioned, this cuts-based analysis will provide the benchmark.

3.1 Data sample and online data-selection

The raw data sample consisted of $\sqrt{s_{NN}} = 13$TeV pp-collisions recorded by the ALICE experiment in 2018. The Monte Carlo productions were produced by the PYTHIA event generator. Two instances of PYTHIA were run, one for producing a prompt-photon dominated signal and one for producing a background with very few prompt photons. In the first case, the events generated by PYTHIA were set to produce at least one prompt or fragmentation photon per event, in so-called γ-jet events. For the background production, jet-jet events are the main events produced, with barely any signal photons.

Cut	Cut Value
Pseudorapidity	$\|\eta\| < 0.67$
Azimuth angle (EMCAL)	$1.4 < \phi < 3.28$
Azimuth angle (DCAL, no hole)	$4.57 < \phi < 5.70$ for $0.23 < \|\eta\| < 0.67$
Azimuth angle (DCAL, no hole)	$5.58 < \phi < 5.70$ for $\|\eta\| < 0.23$
N_{cells} per cluster.	$N^{cells}_{clusters} < 1$
Number of local maxima	$NLM \leq 2$
min distance to bad channel	≥ 2
Shape parameter σ^2_{long}	$0.1 < \sigma^2_{long} < 0.3$
Cluster-prim. track matching	$\|\Delta\eta\| > 0.05$, $\|\Delta\phi\| > 0.05$ and $E_{clus}/p_{track} > 1.75$
Exotics removal	$F_{+} < 0.97$

Table 1: Table providing an overview of the cuts applied to identify EMCal clusters as belonging to photons.

Three different triggering schemes were applied to the data: The minimum-bias (kINT7) trigger, EG1 trigger and EG2 trigger. The efficiency, purity and acceptance was calculated separately for each data sample.

This analysis's "RAW" data has already been through an initial data selection. The specific analysis task was: AliPhysics/PWGGA/GammaConv/AliAnalysisTaskGammaIsoTree.cxx. This task was run on the CERN-computing grid. The main purpose of it was to ensure a good reconstruction quality for the tracks and EMcal-clusters considered in this analysis. Since this task was not developed further as part of this analysis please refer to [1] for details on the online data-selection.

3.2 Signal definition and cuts

This analysis aims to improve the measurement of prompt photons from ALICE data. Analyses like this are bound to be largely dependent on simulations of particle collisions. Therefore, it is important to have a clear signal definition for transparent comparisons to actual data.

Definition 1 *In the simulated particle collisions, a signal photon is defined as an isolated prompt or fragmentation photon.*

The "isolation" requirement requires the energy of clusters surrounding the photon in a cone of a specified size to be less than a given threshold value. For this analysis, the cone size is $R = 0.4$ in terms of η and ϕ and the isolation energy cut is $p_{\mathrm{T}}^{iso} < 1.5\mathrm{GeV}/c$. A more exact definition is given in [1]. Thus, a signal photon must be labelled as a prompt or fragmentation photon and be isolated on the generator level.

Table 1 presents the cuts applied in the baseline analysis. There are also applied cuts in the initial online data selection, which will not be discussed here. The cuts applied on η and ϕ ensure that the clusters and tracks considered here are in the EMCAL and DCAL detector acceptance. The cuts on the number of cells, local maxima and minimum distance to the bad channel are all meant to ensure that only EMCal clusters that are reliably defined are considered. The track-matching cuts are applied to ensure that only clusters from photons are accounted for in the analysis.

3.3 Results of Baseline analysis

The efficiency and purity was determined from the cuts-based analysis from monte-carlo data. The definitions are given here for reference:

$$\varepsilon = \frac{N_{signal}}{N_{signal,gen}} \tag{1}$$

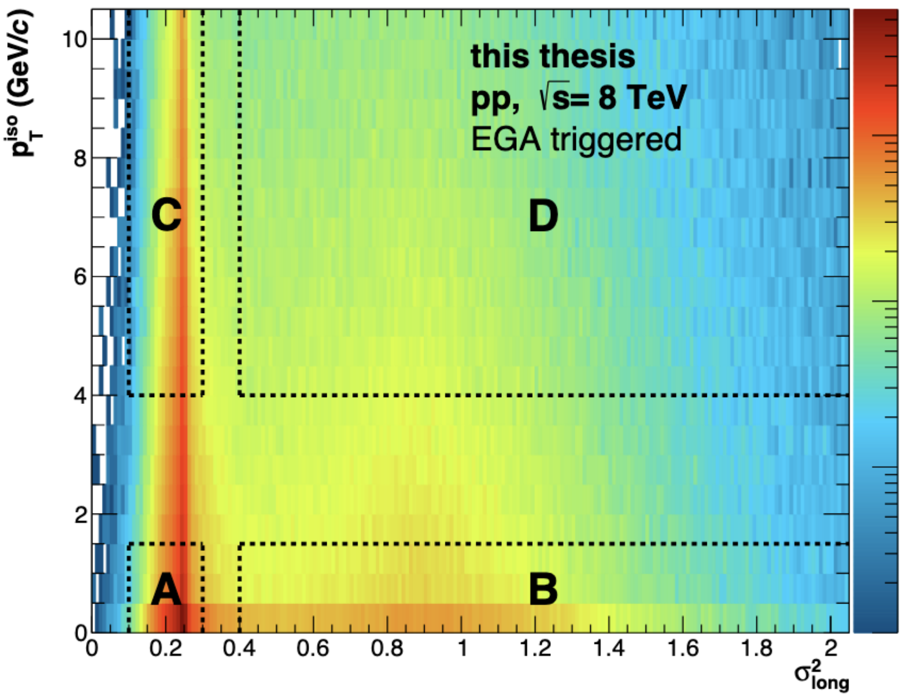

Figure 1: Correlation plot between p_{T}^{iso} and cluster width $M02/\sigma_a$. Notice the highlighted A, B, C and D regions. A corresponds to the signal region. This figure was taken from [1]

$$P = \frac{N_{signal}}{N_{signal} + N_{background}} \tag{2}$$

In addition, a purely data-driven estimate of purity was calculated based on the ABCD method. The idea is to look at the correlation between isolation p_{T}^{iso} and the width of the cluster ($M02$). Clusters from isolated photons are expected to have a low p_{T}^{iso} value and be "narrow". That is, merged π^0 clusters will generally be elliptical. By contrast, a true isolated photon cluster should be circular and have a smaller spacial extent[1]. An example of such a correlation plot is shown in figure 1. Under the right assumptions, the background contribution in the signal region (labelled A) can be constructed from the actual yields in the other background-dominated regions (B,C and D). In that way one can construct an estimate of the purity in the following way[1]:

$$P_{ABCD,raw} \equiv 1 - \alpha \frac{N_C N_B}{N_A N_D} \tag{3}$$

N_C, N_B, N_A and N_D are the number of clusters in the respective regions. α is a correction factor that can be calculated based on MC-predictions. The details of the deriviations and the exact definition of α can be found in [1].

The final results for efficiency and purity (data driven and MC) are shown in figures 2 and 3. The data is minimum bias triggered pp-collisions at $\sqrt{s_{NN}} = 13$TeV. The corresponding results for the EG1 and EG2 data have been relegated to the appendix (section 6). The efficiency was generally lower than expected compared to the $\sqrt{s_{NN}} = 8$TeV results presented in [1]. However, independent measurements of the same efficiency yielded comparable results. These results can be found in [3]. The purity estimates are reasonable compared to the previous results in [1], and it is positive that the two methods are consistent. The fluctuations for P_{abcd} are due to the limited amount of prompt photons in minimum bias-triggered collisions. In short, the results look reasonable, when comparing to other measurements and can provide a benchmark for the machine learning approach.

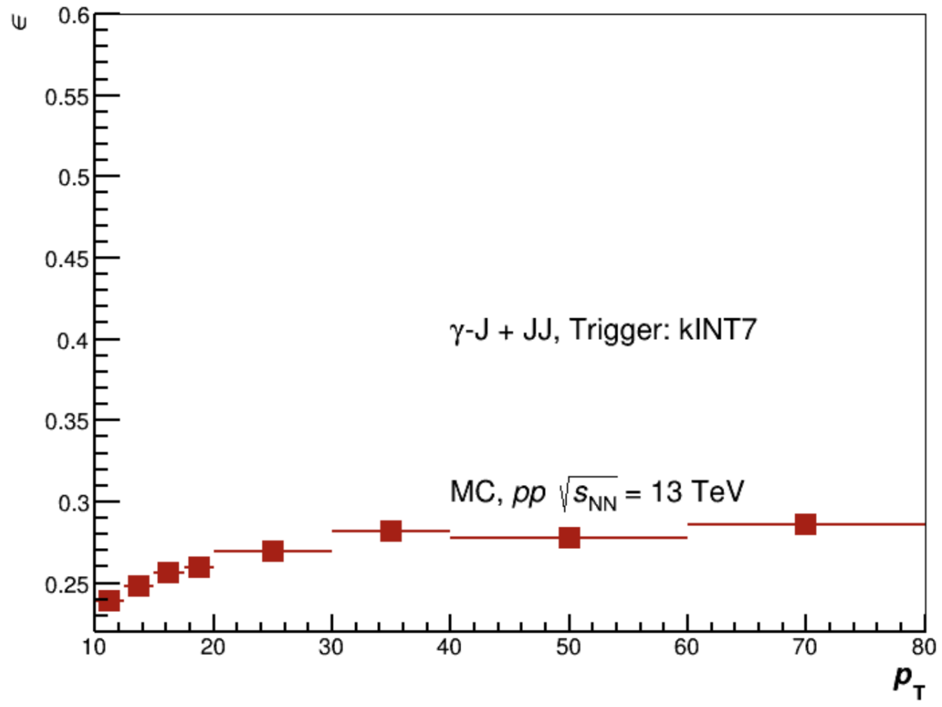

Figure 2:

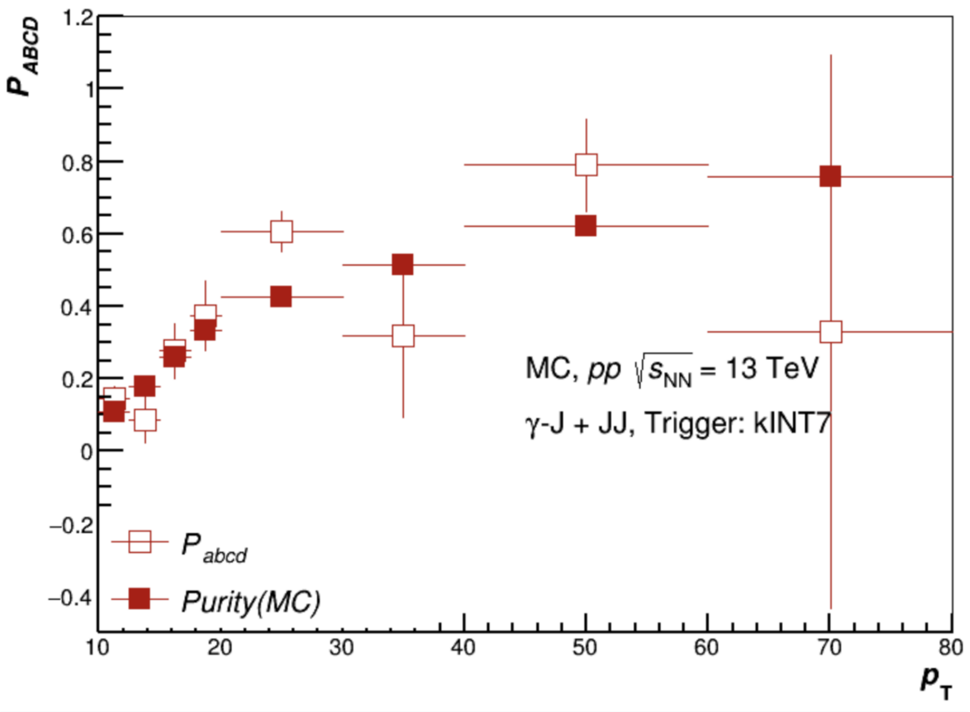

Figure 3:

4 Machine Learning approach

The machine learning approach tested here was a Boosted decision tree algorithm. In a boosted decision tree, a series of simple decision trees of a specified maximum complexity are used to categorise a variable. The complete model is built iteratively. Starting with a default tree. A new tree is then constructed taking into account the efficacy of the previous one. In this way, hundreds of trees are constructed. The final output of the model is a numeric score (a linear combination of the scores from all trees). The algorithm was run via the Hipe4ML-framework[4]. There are two stages to the optimization procedure in Hipe4ML. First, the hyper parameters of the model are optimized using optuna[5]. Then, the actual decision tree is trained by calling the xgboost library[6]. There are other algorithm options that can be called via Hipe4ML, but those features were not explored in detail. For a detailed description of the optimization procedures at the various stages of the training, refer to [4], [6] and [5].

4.1 Data selection and signal definition

The only cuts applied to data were track-macthing and the isolation-cuts. The same cuts were applied as in table 1. This was done to ensure firstly that the model was primarily trained on photons and not charged particles. Secondly, it was done to make sure that the signal definition was comparable to the baseline analysis. For the same reason, the model was not allowed to train on p_{T}^{iso}. The parameters the model was allowed to train on were:

- Number of cells per cluster.
- Number of leading maxima per cluster (NLM).
- Distance to Bad channel.
- The energy fraction carried by the leading cell in cluster(E_{frac}).
- Minimum mass difference to π^0 and η^0.
- Split mass.
- Long ($M02$) and short axis ($M20$) of cluster in terms of η and ϕ.

The minimum mass difference requirement is necessary some photons will be produced by decays from π^0 and η^0. Therefore, one has to check that the energy deposited by the photon is suitably separated from the masses of these particles. The purpose of the split mass is much the same. It is constructed by separating the clusters into two and determining what mass of particle the halphs would correspond to. In this way, one can check if the one cluster has been caused by the decay of a specific particle into two photons. This is assumed to be more likely as p_T increases. The efficiency is calculated from MC data, completely analogous to the cuts-based analysis. For purity, a template fitting procedure was applied. Here, signal and background templates from MC were fitted to the final model output once applied to data.

For testing purposes the signal and background and ”data” came from the same MC-dataset.

4.2 Model setups and results

After a series of initial tests, three models were trained. One, the base line model, was not allowed to train on p_{T}. The second was allowed to train on p_{T} and the third was allowed to train on p_{T} and build deeper trees than the other two. The feature importance can quantify the discriminatory power of each training parameter. This is the purity at each node involving a particular variable, weighted by the number of times this variable appears in the model and the amount of training data that passes through these nodes. To compare the feature importance for different models, the results were scaled by the score of the most important variable in each model. The results for the three main models can be seen in figure 1. For the model not trained on p_{T} the most important parameters are the V1SplitMass, $M20/M02$ and the number of cells per cluster. It is sensible that the V1SplitMass is important, especially at high p_{T}, since the probability of a pion splitting to two photons increases in this regime. Hence, it has good discriminatory power. For $M20/M02$ and the number of cells per cluster, which are all correlated to the shower size and shape. Prompt photons are expected to have relatively narrow showers, which is why the p_{T}^{iso} vs $M02$ correlations could be used to estimate the purity in section 3.3. Therefore, it is reasonable again that the $M20/M02$ and the number of cells per the cluster has strong discriminatory power here. The picture is comparable for the models that were allowed to train on p_{T}. The only difference is that the V1SplitMass is significantly less important due to the correlation with p_{T}.

All in all, the feature importance for all three model looks very reasonable. And one can proceed to calculate efficiency and purities for these results.

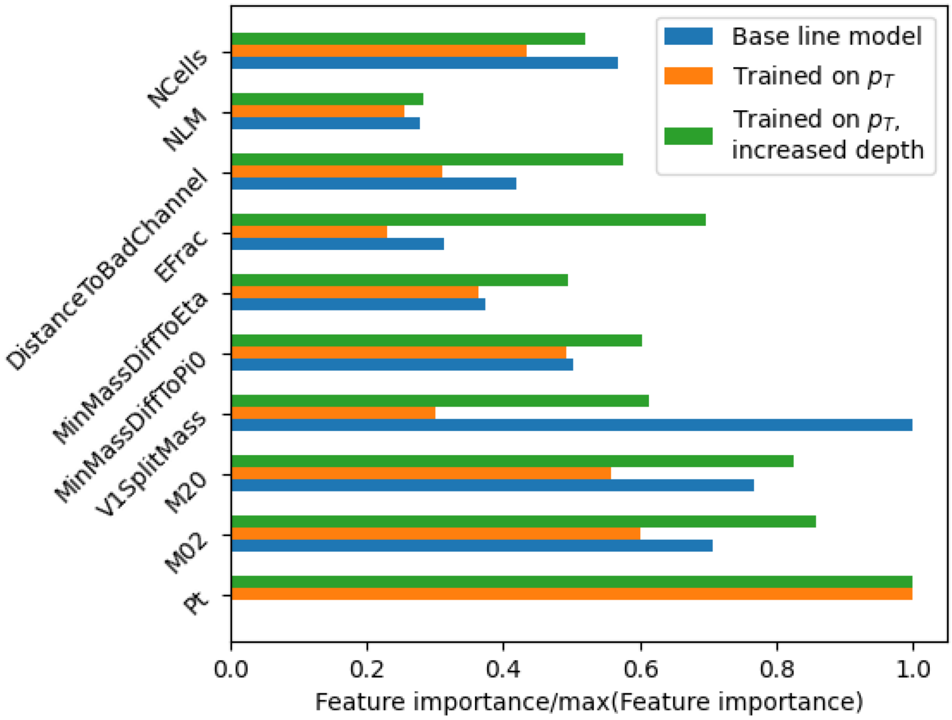

Figure 4: Feature importance for the three main models, the baseline model, a model trained on p_T and a model trained on p_T and allowed to have more complex trees.

4.3 Efficiency and Purity from Template Fits

A template fitting procedure was used to determine the composition of signal and background in a data sample and thus determine efficiency and purity. This method was applied to Monte Carlo data the model had not been trained on. An expression of the following type was fitted:

$$\frac{dN}{dBDT} = A \cdot \left(\frac{dN}{dBDT}\right)_{Signal} + B \cdot \left(\frac{dN}{dBDT}\right)_{Background} \tag{4}$$

Where A and B are fit parameters, $(dN/dBDT)_{Signal}$ is the signal template BDT-distribution and $(dN/dBDT)_{Background}$ is the background template BDT-distribution. The template used in the preliminary tests was the actual signal/background contributions of the Monte Carlo sample. If the template fit was effective, it would reproduce the correct ratio of signal and background. An example of a template fit can be seen in figure 5 the ratio of $Fit/\frac{dN}{dBDT}$. A perfect agreement is found between the actual distribution and the fit, showing that the method works.

The purity can be estimated from the template fit by integrating the BDT distributions of signal and

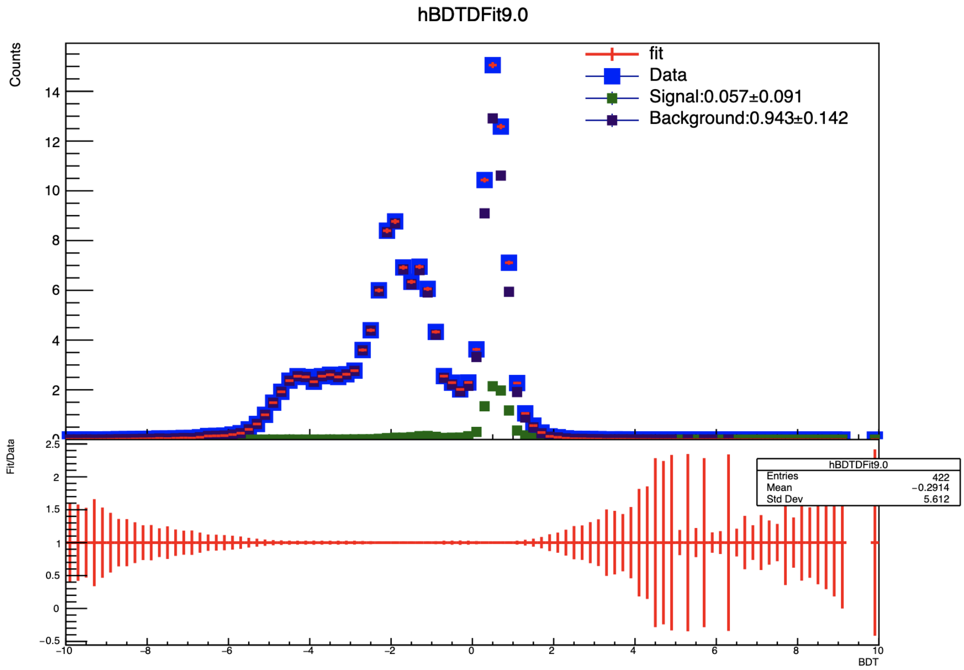

Figure 5: Example of template fit for particles with $9 < p_{\mathrm{T}} < 9.5\mathrm{GeV}/c$. The ratio between fit and true BDT distribution is shown in the bottom plot.

background components scaled by the fit parameters:

$$P_{ML}(p_{\mathrm{T}}) = \frac{A\int_{BDT^a}^{BDT^b}(dN/dBDT)_{Signal}dBDT}{\int_{BDT^a}^{BDT^b}(A\cdot(dN/dBDT)_{Signal} + B\cdot(dN/dBDT)_{Background}dBDT} \tag{5}$$

If one sets the BDT cuts to $\pm\infty$, the purity of the raw sample is recovered. An estimate of the efficiency can also be given. Defined in the following way:

$$\varepsilon_{ML} = \frac{A\int_{BDT^a}^{BDT^b}(dN/dBDT)_{Signal}dBDT}{A\int_{\infty}^{-\infty}(dN/dBDT)_{Signal}dBDT} \tag{6}$$

To get the true efficiency, one has to scale this quantity by the efficiency of the p_{T}^{iso}-cut applied before the machine learning. However that one is quite high, why only the ε_{ML} is recorded here.

One can attempt to apply cuts to the model's BDT values to improve the purity and efficiency. The purity corresponding to the three models under various BDT cuts is plotted in figure 6 and the corresponding efficiencies in figure 7. Generally, the purities reported here are higher compared to the results from the cuts-based analysis. In addition, the efficiency is significantly higher (Even accounting for the missing correction for the p_{T}^{iso}-cut). As for the cuts-based analysis, it is very difficult to get a proper estimate of the purity at very low p_{T}. It seems that one has to apply a lower p_{T} cut here compared to the higher values, and not much is gained by training the model on the p_{T} variable. The efficiency plot also shows that the models trained on p_{T} lose all tracks, going to a low p_{T}, regardless of the cut in the BDT. The behaviour of the photons in this region has to be understood in greater detail to improve this approach. At high p_{T}, the models generally fair very well, and very aggressive cuts on BDT can be applied without impairing the efficiency too much. In short, these are promising first results.

5 Conclusion

Two methods have been tested for identifying prompt photons in pp collisions at $\sqrt{s} = 13\mathrm{TeV}$. One cuts-based analysis forms the baseline, and a machine learning approach employing a boosted decision

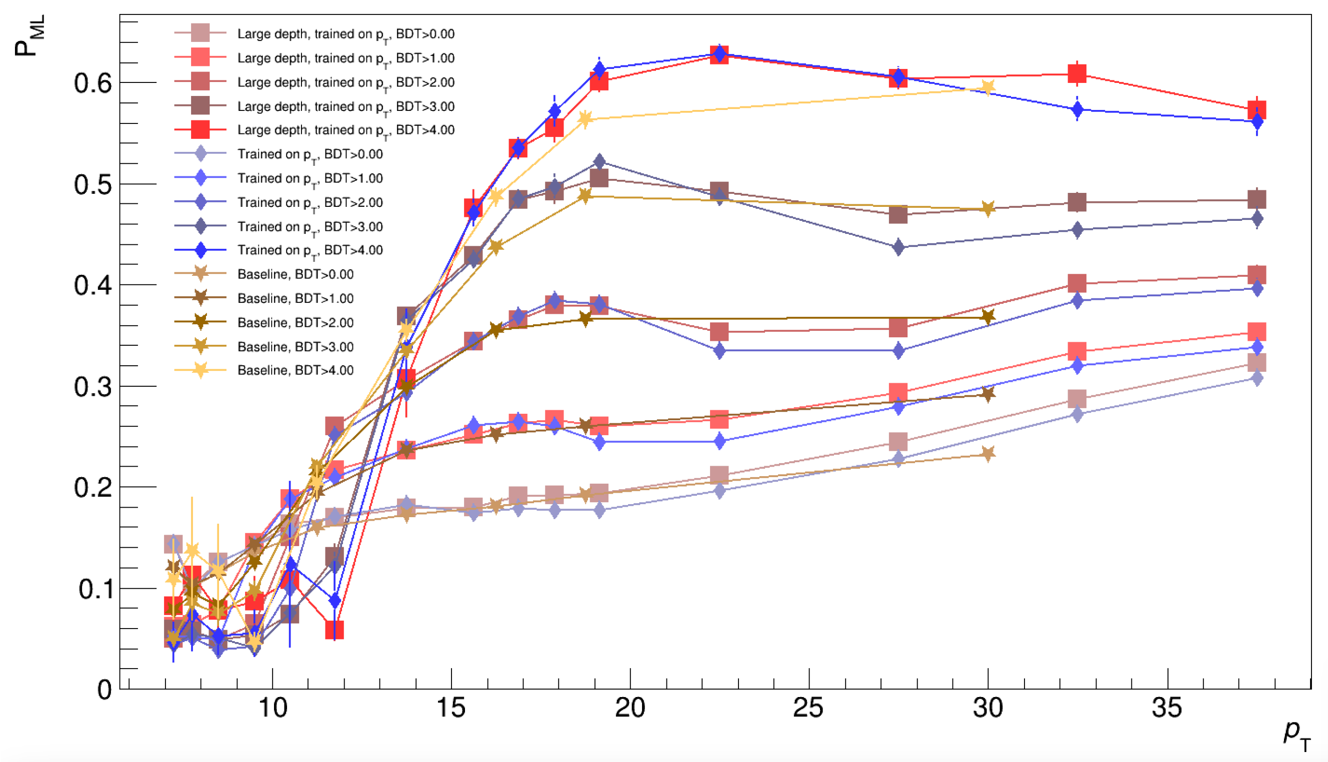

Figure 6: P_{ML} as a function of p_{T} for the three models and various BDT-cuts (min value, going to $+\infty$).

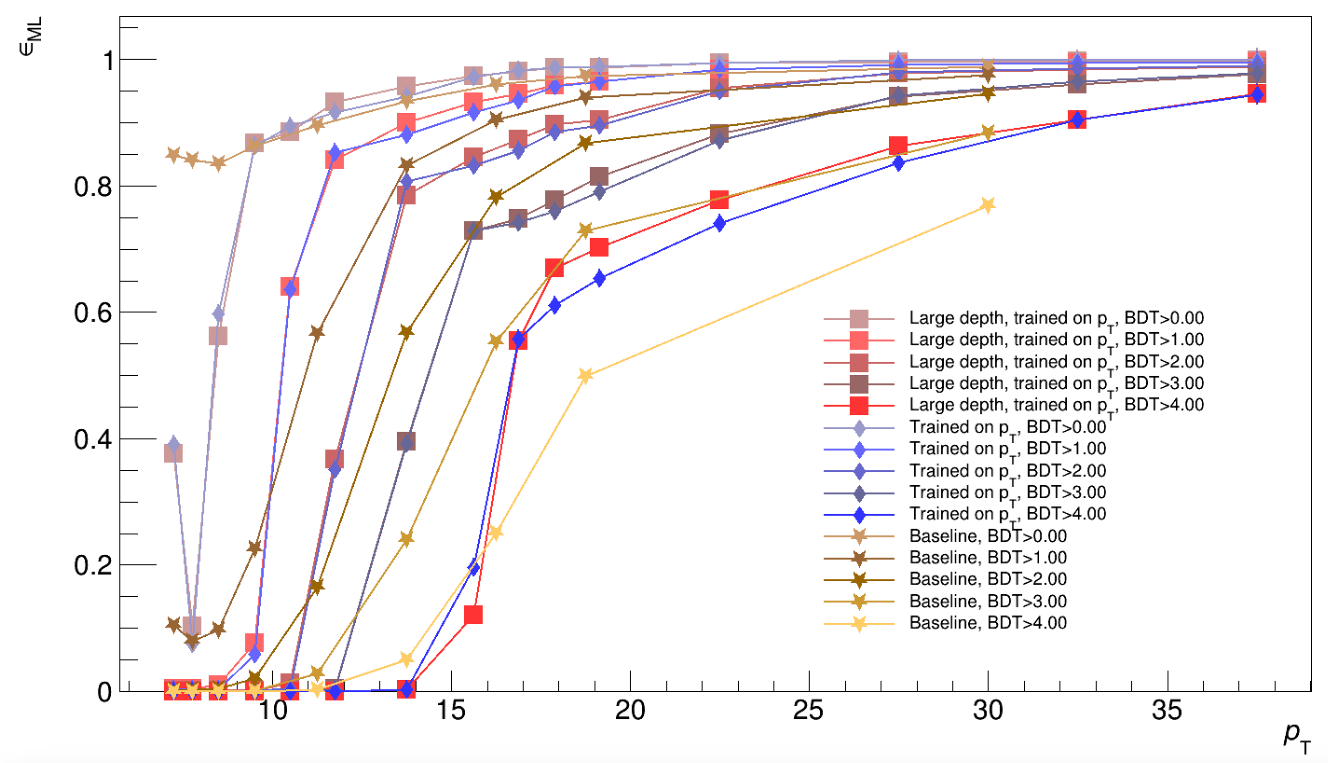

Figure 7: ε_{ML} as a function of p_{T} for the three models and various BDT-cuts (min value, going to $+\infty$).

tree. Preliminary results from the machine learning approach are promising, showing high purity and efficiency. However, further optimisation of the model is needed to probe the low p_{T} part of the spectrum.

188 # 6 Appendix: A1 (Baseline analysis)

189 ## 6.1 Efficiency

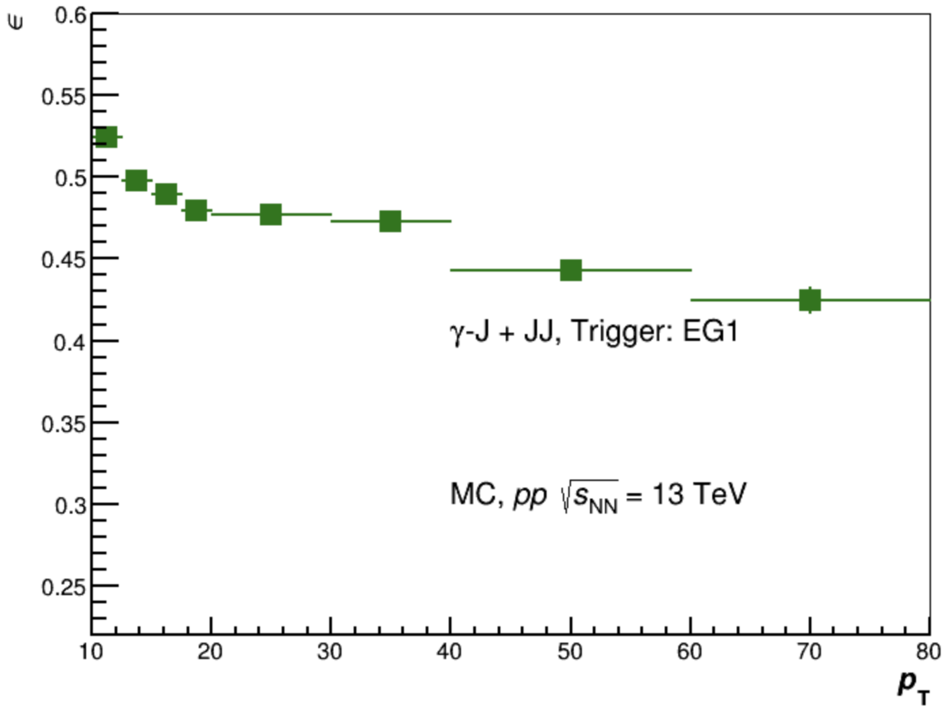

Figure 8: Efficiency for EG1 triggered pp at $\sqrt{s} = 13$TeV.

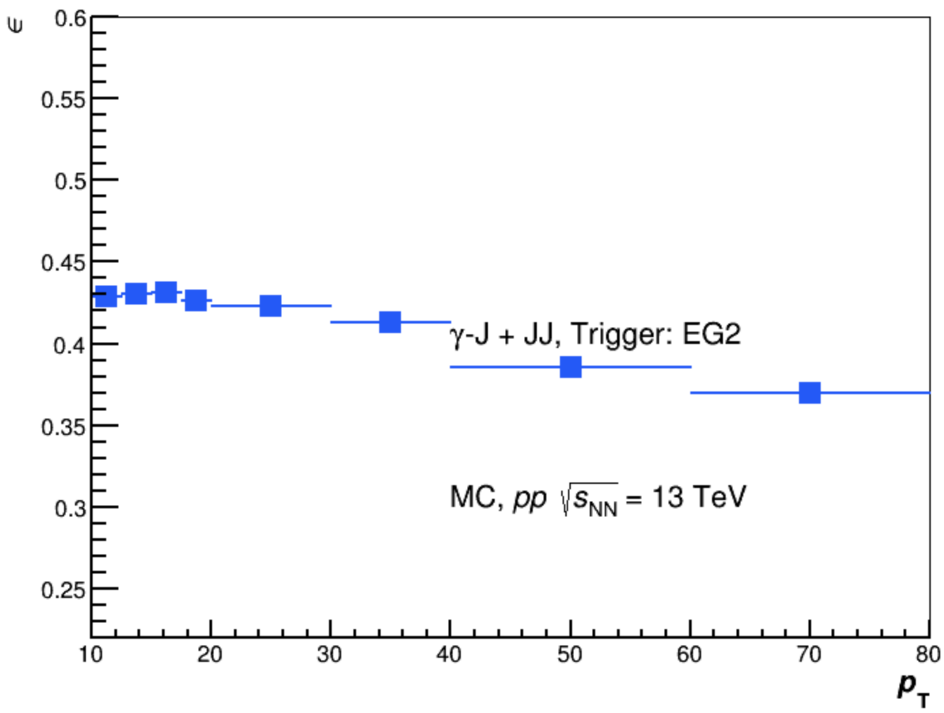

Figure 9: Efficiency for EG2 triggered pp at $\sqrt{s} = 13$TeV.

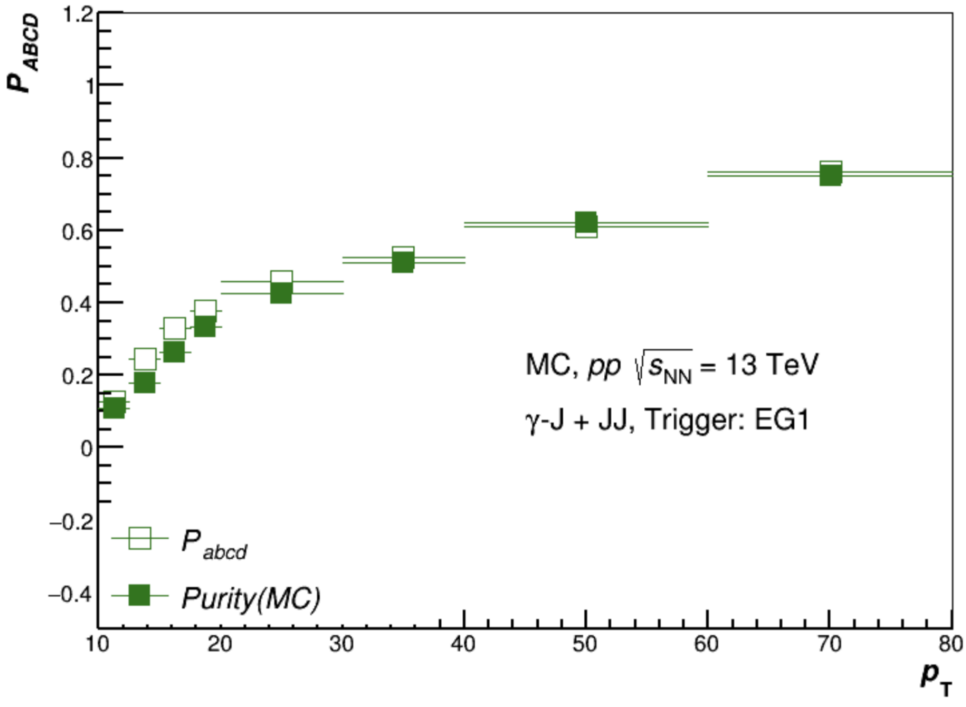

Figure 10: Purity for EG1 triggered pp at $\sqrt{s} = 13$TeV.

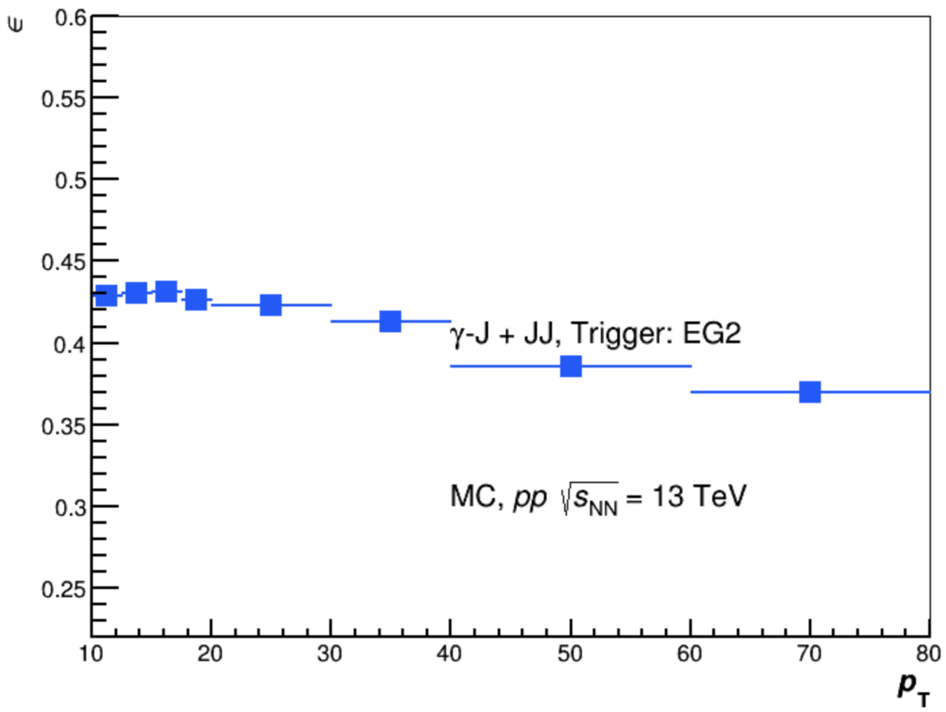

Figure 11: Purity for EG2 triggered pp at $\sqrt{s} = 13$TeV.

6.2 Purity

Acknowledgements

References

[1] F. Jonas, *Probing the initial state of heavy-ion collisions with isolated prompt photons*. PhD thesis, University of California Berkeley, 2023.

[2] ALICE, "Alice technical design reports", 1998-2008.

[3] ALICE, "Measurement of the inclusive isolated-photon production cross section in pp collisions at $\sqrt{s_{NN}} = 13\text{TeV}$",.

[4] Hipe4ML, "Hipe4ml tutorial and documentation", 2024.

[5] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework", in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2019.

[6] C. G. Tianqi Chen, "Xgboost: A scalable tree boosting system", in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794)*. 2016.

205 # A The ALICE Collaboration