ML-based classification of photons and neutral mesons for Pb–Pb collisions recorded by ALICE

Abhishek Nath[1,*] for the ALICE Collaboration

[1]Physikalisches Institut, Ruprecht-Karls-Universität Heidelberg, Heidelberg, Germany

Abstract. Direct photons are unique probes to study and characterize the quark-gluon plasma (QGP) as they are produced at all stages of heavy-ion collisions and leave the collision medium practically unscathed. Measurements at top Large Hadron Collider (LHC) energies at low p_{T} reveal a very small thermal photon signal accompanied by considerable systematic uncertainties. Reduction of such uncertainties, which arise from the π^0 and η measurements and the photon identification, is crucial for obtaining significant results to be compared to the available theoretical calculations.
To address these challenges, a novel approach employing machine learning (ML) techniques has been implemented for the classification of photons. An open-source set of frameworks comprising hipe4ml and scikit-learn packages is chosen for training, validation, and testing the model on a part of Run 2 Pb–Pb data at $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV collision energy.

In this paper, the performance of the novel approach in comparison to the standard cut-based analysis is presented. Initial findings employing gradient-boosted decision trees demonstrate a substantial enhancement in photon purity while preserving efficiency levels comparable to those of the standard cut-based method. Strategies for addressing highly imbalanced data sets, including techniques like feature reduction during training and the implementation of scaled penalty factors to enhance discrimination between signal and background are also addressed. Finally, the feasibility of incorporating such ML methods into the main workflow of direct photon analysis is also presented

1 Introduction

The ALICE [1] (A Large Ion Collider Experiment) experiment is one of the four big experiments at CERN's Large Hadron Collider [2] (LHC). The ALICE experiment aims to explore and characterize the properties of strongly interacting matter under conditions of extreme temperature and energy density. One of the primary objectives of the experiment is to study the quark-gluon plasma [3] (QGP). It is a state of matter in which quarks and gluons are no longer confined within hadrons. This phase is believed to have existed in the early universe, up to a few microseconds after the Big Bang. By colliding heavy ions at very high energies, ALICE aims to recreate and investigate the QGP and thereby study and inspect fundamental

*e-mail: aabhishek.naath@gmail.com

 https://doi.org/10.1051/epjconf/202533701137

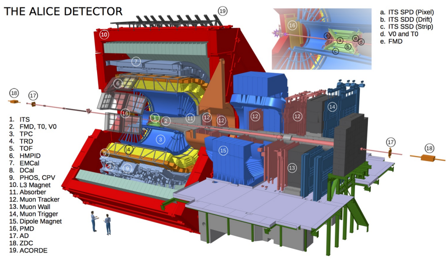


Figure 1. ALICE detector schematic for Run 2

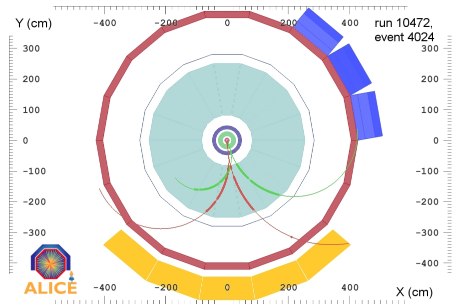


Figure 2. Photon conversion Method: Sketch of how the secondary vertex (π^0) is identified via charged tracks

aspects of Quantum Chromodynamics (QCD), namely color deconfinement, chiral symmetry restoration, and the properties of the hot QCD medium.

There are many probes suitable for studying the QGP in heavy-ion collision. However, being colorless, direct photons [4] stand out as one of the best candidates. The photons do not interact strongly with the medium and thus carry undistorted information about their production throughout the evolution of the heavy-ion collision. Further, photons are produced at all stages of heavy-ion collision, providing an integrated picture of the evolution and dynamics of QGP.

The ALICE detector system consists of several sub-detectors. Figure 1 shows the ALICE setup in Run 2. The ALICE detector is particularly well-suited for direct photon measurements at low-p_T (a few MeV/c) regions due to the ITS [5] and the TPC [6] which provide excellent tracking capabilities in the central barrel. Both the ITS and TPC can be used to measure photon yields by reconstructing track pairs originating from pair production in the detector materials. This approach offers high resolution of the transverse momentum (p_T) of photons, particularly for low-p_T photons (ranging from a few hundred MeV/c). For high p_T photons (in the range of tens of GeV/c), the calorimeters provide great momentum resolution. Such high-resolution measurements of photons across a large p_T range provide a great starting point for detailed studies on direct photon yields, azimuthal anisotropies (e.g., elliptic flow v_2), and their dependence on collision systems and energy [7, 8].

With the advent of better computing hardware and infrastructure, the usage of machine learning (ML) in data analysis has become quite common in the scientific communities. In recent years, ML methods have been applied in various ways in ALICE. Usage of classical machine learning as well as deep learning in particle identification, reconstruction, and even in event generation has been studied and implemented [9].

2 Photons

2.1 Photons Conversion Method

The photon conversion method (PCM) in ALICE is used to reconstruct photons by detecting tracks of electron-positron (e^+e^-) pairs which are generated by pair production in the presence of the detector material (namely, the beam pipe, the ITS, the TPC and their support structures). This is shown in Figure 2 [10]. The tracking of charged particles is done using the TPC and the ITS. The photon is separated from the dataset by calculating the invariant mass of the pair, which is near zero in genuine conversions, and by applying kinematic cuts like pair opening angle as well as electron identification via specific energy loss $\mathrm{d}E/\mathrm{d}x$ in the TPC. More details about the process can be found in [11].

2.2 Photons as a probe

Direct photons are those that do not originate from the decay of hadrons. They are produced at various phases of a collision, such as initial hard scattering (e.g. $q + g \rightarrow q + \gamma$ and $q + \bar{q} \rightarrow g + \gamma$), thermal radiation of the QGP, collision in the hadron gas, and jet-medium interactions. The findings of WA98 [12], PHENIX [13], ALICE [14], and STAR [15] validate a component of thermal origin, especially for low p_{T}. Direct photon (v_2) [16], reveal medium features. Decay photons, primarily from hadron decays ($\pi^0 \rightarrow \gamma\gamma$ and $\eta \rightarrow \gamma\gamma$), constitute the dominating background for direct photon measurements. Additionally, Dalitz and secondary decay photons (i.e. photons from decays of secondary hadrons) are also present in the background. Direct photons are distinguished experimentally by determining the inclusive-to-decay photon ratio.

Previous results on the Run 2 dataset with standard cut-based analysis yielded an inclusive-to-decay photon ratio (R_γ) not significantly different from unity in the low-p_{T} region [17]. This is the region of interest from a physics perspective to detect the thermal photon signal and the previous analysis reveals that the signal is too low and is accompanied by large uncertainties to be significant specifically in this region. Hence improving the systematic uncertainties here deserves special attention. In this regard, including machine learning models is considered. In this paper, we demonstrate the first result that we obtained and not yet the full study of systematic uncertainties.

3 XGBoost

XGBoost[18] (eXtreme Gradient Boosting) is a highly efficient, scalable, and widely used machine learning algorithm based on the gradient boosting framework. Gradient boosting combines predictions from multiple decision trees, constructed iteratively, to minimize a loss function and improve predictive accuracy. XGBoost enhances this framework with features such as regularization to prevent overfitting, parallel processing for faster training, and optimized handling of missing data. It is particularly well-suited for tabular datasets with structured features, providing state-of-the-art performance in regression, classification, and ranking tasks [19, 20]. In high-energy physics, XGBoost is commonly used for tasks like event classification, signal and background discrimination, and feature selection due to its interpretability, efficiency, and robustness against noise [21].

4 Analysis

4.1 Data selection

To create a p_{T}-dependant photon XGBoost model, 16 p_{T} intervals were selected and the models were trained in these p_{T} intervals. For training and validating the photon XGBoost model, Monte-Carlo (MC) generated data were used. The MC generated data were produced via the AliRoot [22] software. Three sets of MC datasets were used as specified in Table 1. Initially, a ROOT TTree file is created with each feature as one of its branches using the Lego train framework [23]. The entries in each branch were synced with the event. The training is performed with these ROOT files with TTrees taken as input. The training is

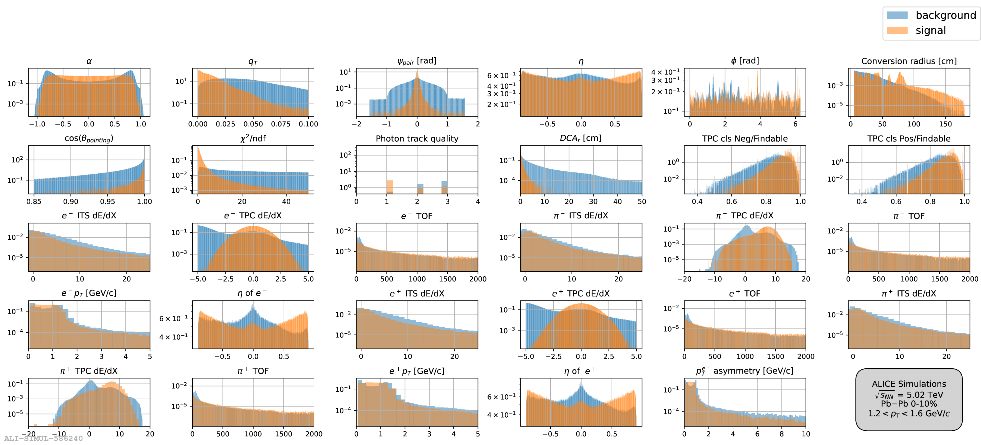


Figure 3. All features corresponding to the photon training for the ALICE Pb–Pb 2018 MC central collision dataset. The units of the x-axis are arbitrary or mentioned in the heading. The units of the y-axis are density (i.e., not pure count). The blue entries correspond to the background data, while the orange is the signal, both of which are verified with Monte-Carlo (MC) truth.

Table 1. Run 2 datasets used for training and testing. $N_{\rm sig}$ indicates the total number of primary converted photons (γ), while N_{bkg} represents the total number of background candidates in the $p_{\rm T}$ range 0–14.0 GeV/c for central collisions (0-10%) only.

Dataset	Generator	$N_{\rm sig}$	N_{bkg}
LHC20e3a	HIJING Minimum Bias	6 M	62 M
LHC20e3b	HIJING 0–10%	4.8 M	50 M
LHC24a1	HIJING + embedded γ (custom $p_{\rm T}$)	77.6 M	14.1 M

done with hipe4ml [24], which is a wrapper package of Scikit-learn focused towards high-energy physics. The models, once trained over all $p_{\rm T}$ ranges, were compiled and deployed via Treelite software in the AliRoot framework.

Our signal features are comprised of MC charged track properties. True signal entries are those track pairs that have a common mother particle (photon) coming from the primary collision. The features are listed in the Figure 3. All of these features were used in the training. Some of the most important features which we can already see from Figure 3 are: The n_σ values of TPC dE/dx for both positive and negative tracks, $q_{\rm T}$ (normalized $p_{\rm T}$ component in the plane perpendicular to the beam), ψ_{Pair} (angle between the plane of the $e^+e^-$of the track pair and the transverse plane), cosine of the pointing angle between the momentum of the photon candidate and the line joining photon candidate to primary collision vertex, track quality (whether the track is present entirely in ITS, TPC or both), χ^2/ndf (quality of track fits for photon), DCA (distance of closest approach of the e^+e^-tracks associated with the photon candidates) etc.

4.2 Model training

The training of the model is done by tuning the XGBoost hyperparameters. Optuna [25] is used for this purpose, and the model is tuned over the given range of hyperparameters shown in Table 2. The cross-validation scoring was done using the ROC-AUC curve, with 5 trials

Table 2. Hyperparameter ranges for XGBoost model optimization for each p_T interval. For our case, i.e., photon training, we only included the first seven of the eight hyperparameters listed below.

Hyperparameter	Range
learning_rate	(0.001, 0.9)
gamma	(0.001, 0.1)
max_depth	(2, 10)
min_child_weight	(2, 10)
n_estimators	(30, 100)
subsample	(0.3, 1)
colsample_bytree	(0.1, 1.0)
scale_pos_weight	(bkg/sig * 0.9, bkg/sig * 1.1)

and a timeout value of 60 sec. The test-to-training ratio was 3:1, and the cross-validation was done with 5 subsets. If we consider our primary datasets: LHC20e3a and LHC20e3b from Table 1, we see that it is highly imbalanced. Hence, to get more converted photon statistics, we introduce a new signal-enriched dataset called LHC24a1. Our final dataset is prepared by taking all these MCs together for central collisions (0-10%) only. We train the models in each p_T bin, optimizing them for each p_T interval by tuning some of the XGBoost model hyperparameters as mentioned in Table 2.

We also considered exploring the identification of mesons (signal pairs corresponding to $\pi^0 \to \gamma\gamma$ and $\eta \to \gamma\gamma$ decay channel) from converted photons as a probable next step of our analysis. Here, the signal-to-background ratio reached over 1000 in some p_T intervals even after including the signal-enriched datasets. Only for such cases, we chose to fine-tune the "scale-pos-weight" hyperparameter (which controls the balance of positive and negative weights) close to the ratio of background to signal in each p_T bin. However, we discontinued this path as of now to focus entirely on photon training, where we exclude the scale-pos-weight hyperparameter.

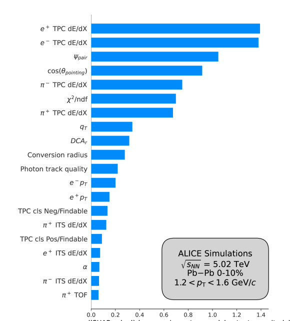


Figure 4. Post-training feature importance plot shows that the TPC electron identification via dE/dx for the charged tracks is the dominant feature for the central collision dataset.

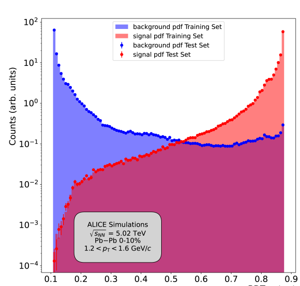


Figure 5. Test [dots] and training [solid] distribution. For both signal and background, they overlap implying a good training result without overfitting.

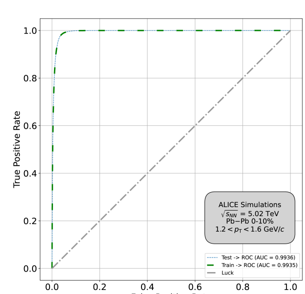
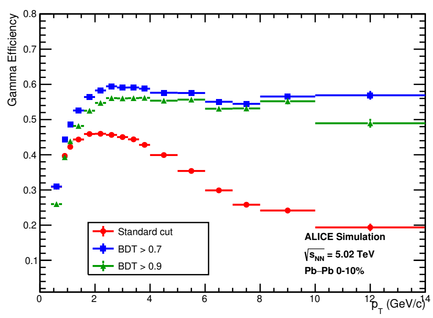


Figure 6. ROC curve showing both training and testing far from the $y = x$ line implies most of the data which are correctly classified and the chance of false classification is very low.

4.3 Model testing and feature importance

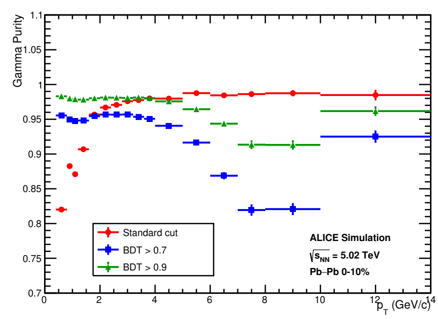
The post-training feature importance plot shown in Figure 4 reveals the most relevant observables from the feature space plot. The feature importance is calculated via the SHAP values [26]. SHAP values are based on Shapley values from cooperative game theory, assigning each feature a contribution to the final prediction. The feature importance is calculated using the mean absolute SHAP value, ensuring a consistent ranking of features. Figure 5 presents the training and test set distribution for background and signal as a function of the boosted decision tree output variable. For both signal and background, the test distribution closely follows the training distribution implying no overfitting. Further, the ROC plot shown in Figure 6 implies that the model's training is both highly accurate and precise.

5 Results



ALI-SIMUL-586280

Figure 7. Photon efficiency as a function of p_T. Both the XGBoost models gives excellent efficiency. From 1 GeV/c, it surpasses the efficiency of the standard cut-based analysis significantly.



ALI-SIMUL-586275

Figure 8. Photon purity as a function of p_T. It is high at low p_T for the XGBoost models at different cutoffs. However, with an increase of p_T there is a drop in purity improved with the higher BDT cutoff model. This suggests implementing a dynamic BDT cutoff model over the p_T range

In Figure 7 the efficiency [1] of the standard cut-based analysis is compared with the XGBoost models at different BDT cutoff values. Here the 'standard cut' refers to the photon classification done by manually selecting cut values for different features based on physical or topological aspects. The BDT model efficiencies are quite high compared to the standard cut efficiency for most of the entire p_T range. In Figure 8, the purity [2] is compared between the standard analysis and the XGBoost models. Here we see that in the low p_T region, the purity is significantly larger than the standard cut based purity. For a model with a BDT score cutoff of 0.9, we achieve over 90% purity. At higher photon p_T the purity of the XGBoost selected sample drops, but less so for the higher BDT cut values. This suggests a p_T-dependent BDT cut to optimize between efficiency and purity.

The results from the photon efficiency studies can be transferred to the $\pi^0 \rightarrow \gamma\gamma$ analysis by using the same selection criteria for the decay photons as described earlier. Photons

[1] all XGBoost true photons/all converted photons
[2] all XGBoost true photons/all XGBoost selected photons

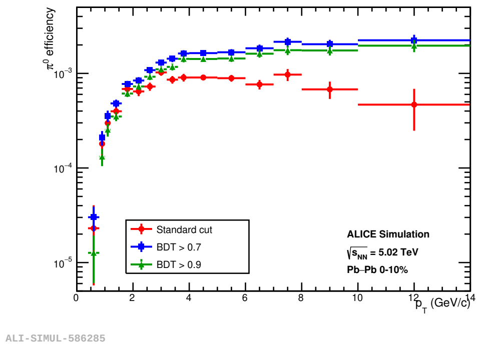


Figure 9. The π^0 efficiency for the XGBoost models compared with the standard cut-based analysis. The efficiency remains significantly higher in $p_T > 2$ GeV/c and comparable at low p_T.

identified by XGBoost, using the MC truth information, can be used to validate photon pairs originating from the same neutral pion ($\pi^0 \to \gamma\gamma$). The π^0 efficiency is then calculated as the ratio of the total number of validated and reconstructed photon pairs originating from the same π^0 in the acceptance region ($|\eta| < 0.8$) to the total number of validated true primary π^0 mesons that are generated with their daughter photons within the same acceptance region ($|\eta| < 0.8$). Figure 9 shows the π^0 efficiency with converted photons selected by the XGBoost model, which exhibits a remarkable gain across a broad p_T range (from 2 GeV/c) and above) compared to the standard method.

6 Conclusion

In this study we explored the use of machine learning, specifically extreme gradient-boosted decision trees (XGBoost), to enhance photon identification in ALICE Pb–Pb central collisions at $\sqrt{s_{NN}} = 5.02$ TeV. By comparing the ML-based approach with the conventional cut-based method, we demonstrated in the most challenging region below 2 GeV/c a significant improvement in photon purity while maintaining comparable efficiency levels. At higher p_T a very high efficiency can be maintained at some expense in purity. An optimum will be sought by p_T-dependant BDT cut values. The results indicate that ML techniques offer a promising avenue for reducing systematic uncertainties associated with direct photon measurements. Future work will focus on further optimization of the ML models and thereby looking for an optimum in systematic uncertainties in particular in the low-p_T region (a few GeV/c). Further, XGBoost models for neutral mesons (π^0 and η) selection are also being explored.

References

[1] K. Aamodt et al. (ALICE), "The ALICE experiment at the CERN LHC" *Journal of Instrumentation* **3**, S08002 (2008)

[2] L. Evans and P. Bryant, "LHC Machine", *Journal of Instrumentation*, **3**, S08001 (2008)

[3] S. Acharya et al. (ALICE), "The ALICE experiment: a journey through QCD", *European Physical Journal C* **84**, 813 (2024)

[4] G. David, "Direct real photons in relativistic heavy ion collisions", *Reports on Progress in Physics*, **83**, 046301 (2020)

[5] M. Faggin, "Vertexing detectors and vertexing performance in Run 2 in ALICE." arXiv:2008.13530 [physics.ins-det] (2020)

[6] J. Alme et al. (ALICE), "The ALICE TPC, a large 3-dimensional tracking device with fast readout for ultra-high multiplicity events." *Nuclear Instruments and Methods in Physics Research Section A*, **622**, 316-367 (2010)

[7] S. Acharya et al. (ALICE), "Direct photon elliptic flow in Pb-Pb collisions at $\sqrt{s_{NN}}$ = 2.76 TeV", arXiv:1805.04403 [nucl-ex] (2018)

[8] V. Pacík et al. (ALICE), "Elliptic flow of identified hadrons in small collisional systems measured with ALICE," *Nuclear Physics A*, **982**, 451-454, (2019)

[9] Ł. K. Graczykowski et al. (ALICE), "Using Machine Learning for Particle Identification in ALICE," arXiv:2204.06900 [nucl-ex] (2022)

[10] K. Koch et al. (ALICE), "π^0 and η measurement with photon conversions in ALICE in proton-proton collisions at $\sqrt{s}$ = 7 TeV", *Nuclear Physics A*, **855**, 281-284 (2011)

[11] S. Acharya et al. (ALICE), "Data-driven precision determination of the material budget in ALICE," *Journal of Instrumentation*, **18**, P11032 (2023)

[12] M. M. Aggarwal et al. (WA98 Collaboration), "Observation of Direct Photons in Central 158A GeV Pb^{208} + Pb^{208} Collisions," *Phys. Rev. Lett.*, **85**, 3595-3599 (2000)

[13] A. Adare et al. (PHENIX Collaboration), "Observation of Direct-Photon Collective Flow in Au+Au Collisions at $\sqrt{s_{NN}}$ = 200 GeV," *Phys. Rev. Lett.*, **109**, 122302 (2012)

[14] J. Adam et al. (ALICE Collaboration), "Direct photon production in Pb–Pb collisions at $\sqrt{s_{\mathrm{NN}}}$ = 2.76 TeV," *Phys. Lett. B* **754**, 235-248 (2016).

[15] STAR Collaboration, "Direct virtual photon production in Au+Au collisions at $\sqrt{s_{\mathrm{NN}}}$ = 200 GeV," *Phys. Lett. B* **770**, 451-458 (2017)

[16] S. Acharya et al. (ALICE), "Direct photon elliptic flow in Pb–Pb collisions at $\sqrt{s_{\mathrm{NN}}}$ = 2.76 TeV," *Phys. Lett. B* **789**, 308-322 (2019)

[17] M.C. Danisch, "Direct Photon Production and HBT Correlations in Pb–Pb Collisions at $\sqrt{s_{\mathrm{NN}}}$ = 5.02 TeV with the ALICE Experiment," *Acta Phys. Pol. B Proc. Suppl.* **16**, 1, A122 (2023)

[18] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System", *arXiv:1603.02754 [cs.LG]*, 2016, https://github.com/dmlc/xgboost.

[19] S.S. Dhaliwalet al., "Effective Intrusion Detection System Using XGBoost," *Information*, **vol 9**, no. 7, p. 149, 2018, https://doi.org/10.3390/info9070149.

[20] W. Chang et al. , "A Machine-Learning-Based Prediction Method for Hypertension Outcomes Based on Medical Data," *Diagnostics*, vol. 9, no. 4, p. 178, 2019, https://doi.org/10.3390/diagnostics9040178.

[21] V. Vatellis, "Advancing Physics Data Analysis through Machine Learning and Physics-Informed Neural Networks," arXiv:2410.14760 [hep-ph], 2024

[22] R. Brun et al., "Computing in ALICE," *Nucl. Instrum. Methods Phys. Res. A* **502**, 339-346 (2003), https://github.com/alisw/AliRoot.

[23] M. Zimmermann et al. (ALICE), "The ALICE analysis train system," *J. Phys.: Conf. Ser.* **608**, 012019 (2015)

[24] L. Barioglio, et al., "hipe4ml/hipe4ml" *Zenodo* (2021), https://doi.org/10.5281/zenodo.5070131.

[25] T. Akiba, et al. "Optuna: A Next-generation Hyperparameter Optimization Framework." arXiv preprint arXiv:1907.10902, 2019, https://github.com/optuna/optuna

[26] S. M. Lundberg, et al., "A Unified Approach to Interpreting Model Predictions", arXiv:1705.07874 (2017)