PROCEEDINGS OF SCIENCE

PoS(ICRC2023)191

Towards searching for ultra-high-energy photons with deep learning techniques

Eleonora Guido,[a,*] **Marcus Niechciol**[a] **and Markus Risse**[a]

[a] *Center for Particle Physics Siegen, Department of Physics, University of Siegen*
Walter-Flex-Str. 3, 57072 Siegen, Germany

E-mail: guido@hep.physik.uni-siegen.de

In the last few decades, it has been proven that ground-based experiments devoted to the study of ultra-high-energy (UHE) cosmic rays are also powerful tools in the quest for UHE photons. The search for these elusive particles, never unambiguously detected so far, relies on the capability of distinguishing air showers initiated by primary photons among the background of the ones generated by nuclei. In this study, we explore the possibility of exploiting an array of water-Cherenkov detectors to distinguish photon-induced air showers based on the shape of the signal traces recorded by the individual detector stations of the array, using the experimental setup of the Pierre Auger Observatory [1] as an example. A photon-induced air shower is dominated by its electromagnetic component, which tends to reach the ground later and to be more spread in time than the muonic one. Maximizing the discrimination power by considering the whole time evolution of the signals implies dealing with hundreds of variables. Additionally, we have to take into account other dependences of the signal shape, such as the energy and zenith angle of the primary particle and the distance of each station from the shower core. For such reasons, we make use of deep neural networks. Here we explore a Convolutional Neural Network algorithm and test it on air shower simulations. We show that, thanks to this innovative approach, it is possible to reach high levels of accuracy in classifying simulated air shower events, providing a promising tool to distinguish UHE photons.

The 38th International Cosmic Ray Conference (ICRC2023)
26 July – 3 August, 2023
Nagoya, Japan

ICRC2023
38th International Cosmic Ray Conference
The Astroparticle Physics Conference

[*] Speaker

© Copyright owned by the author(s) under the terms of the Creative Commons
Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0).

1. Introduction

Cosmic rays with energies $\gtrsim 10^{18}$ eV are generally referred to as ultra-high-energy (UHE) cosmic rays and can be observed through the extensive air showers (EAS) they produce when interacting with air nuclei in the atmosphere. Although they have been studied for more than 60 years with increasingly larger detector arrays, many questions about their nature and origin are still open.

Before reaching the Earth's atmosphere, UHE cosmic rays may interact with background radiation photons, both in the source ambient and during their propagation, resulting in the production of UHE photons. Since the flux of such photons is expected to be related to the composition and the energy distribution of the parent nuclei at the sources and to the interaction processes they undergo in the intergalactic medium, their observation can set constraints on the models of origin and propagation of cosmic rays.

Both photons and nuclei at UHE generate EAS in the atmosphere, which can be observed by large ground-based experiments, for example exploiting arrays of water-Cherenkov detectors. A photon-induced air shower has a strong electromagnetic and a negligible hadronic component, which impacts both the particle content and the development of the shower in the atmosphere. No UHE photons have been unambigously detected so far [2], mainly because of the low rate of particles at UHE and the limitations in the capabilities to distinguish the showers generated by primary photons from the ones initiated by protons or heavier nuclei.

In this analysis we explore the possibility of distinguishing photon-induced showers by using the shape of the signal traces recorded in the individual stations of an array of water-Cherenkov detectors, using the experimental setup of the Pierre Auger Observatory [1] as an example. Considering that the electromagnetic component tends to arrive later and to be more spread in time with respect to the muonic one, the shape of a signal trace produced by a photon-generated air shower exhibits a broader shape with respect to a proton-induced one, the energy and the zenith angle being equal.[1] As an example, in Fig. 1, the cumulative functions of the traces for the first 150 recorded time bins are shown in the case of particles with similar energy and zenith angle, as detected in the detector station measuring the highest signal of the event.

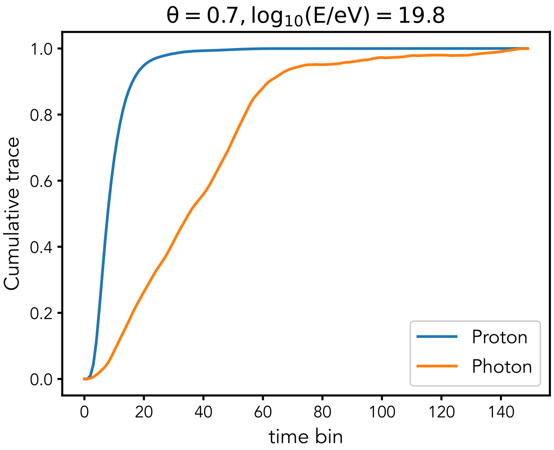

Figure 1: Simulations of cumulative time traces as recorded with a water-Cherenkov station, in the case of a proton and a photon with similar energy and zenith angle.

[1] Note that eventually we will aim at distinguishing photons from nuclear species in a sample of real events, where also nuclei heavier than protons are observed. However, since protons generate the air showers that can be most likely misinterpreted as photons, here we choose the most conservative approach by only comparing photon-induced air showers with the ones generated by protons.

PoS(ICRC2023)191

One of the novelties of this approach with respect to previous analyses dealing with SD traces (e.g. [3, 4]) lies in the fact that we consider the whole time evolution of the signal instead of exploiting for example the risetime observable, which is related only to its very first portion. In fact, it is clear from Fig. 1 that the signals exhibit distinctive features both in their rising and falling parts, so that the discrimination power can be enhanced by extending the considered time range. Besides, the signal shape exhibits also dependencies on other quantities that have to be taken into account, such as the energy and the zenith angle of the primary particle and the station distance from the shower core.

Exploiting machine learning techniques is the most suitable choice to deal with hundreds of variables and identify complex patterns in the data set. Several classical algorithms and deep learning ones have been tested, achieving comparable discrimination power. For the analysis presented here, we choose to focus on the results obtained with a Convolutional Neural Network (CNN), which is the most promising one in anticipation of further extending this analysis, for example including more stations and considering their footprint on the surface detector array.

PoS(ICRC2023)191

2. Simulated events and input information

CORSIKA [5] simulated air showers are generated with EPOS-LHC [6] and FLUKA [7] as the hadronic interaction models at high energy and low energy, respectively. Since we choose the Surface Detector (SD) of the Pierre Auger Observatory as a case study, we simulate the detector response to air showers provided in the same format as real events with the Auger Offline package [8]. The set of simulated events consists of ∼ 92 000 photons and ∼ 30 000 protons, with energies between $10^{18.5}$ and $10^{20.5}$ eV and zenith angles $\theta < 60°$. The main idea consists of providing our network the raw reconstructed information of simulated events as input, focusing on the three triggered SD stations which record the highest signals for each event.

The following quantities at the level of individual station are used as input variables:

Signal trace In each station, the total signal trace is provided by three PMTs through six flash analog-to-digital converters (FADCs). It consists of 768 time steps with a length of 25 ns each. Here we consider as input variables the cumulative functions of the traces for the first 150 time bins, which corresponds to a time window of 3.75 µs.[2] By construction, the values range between 0 and 1 and do not need further re-scaling.

Distance from the shower core. In general, the closer the station is to the shower core, the larger the signal. The distance of each considered station is re-scaled by expressing it in units of the SD array spacing (1500 m) and by being subtracted by the average of all the distances in the whole data sample. In such a way, the input distances $\tilde{d}$ cross over positive and negative values and never exceed few units.

Total signal. The total signal recorded in the stations S_{tot} is expressed in vertical equivalent muons (VEM). It is related to the energy of the primary particle and computed integrating over a time range longer than the one considered for the traces. It is re-scaled as:

$$S_{\text{tot}}^{\text{rescaled}} = \frac{\log_{10}(S_{\text{tot}} + 1)}{\log_{10}(S_{\text{norm}} + 1)},$$

[2] We have also verified that increasing the number of bins up to 300 does not significantly improve our discrimination power, since most of the information is provided in the rising part of the signal.

where $S_{\mathrm{norm}} = 100$ VEM. The rescaled values are always of the order of unity.

Azimuth in shower plane coordinates. It is given by the projection of the position of the station on the shower plane with respect to the direction given by the projection of the shower axis on the ground. Since early-triggered detectors exhibit smaller angles, it allows the reconstruction of the shower geometry, along with the distance from the shower core. It is expressed in radians.

In addition to the station-related variables some information at the level of the event has to be considered, as well:

Zenith angle. The shape of the recorded signals depends on the inclination of the showers, given by the zenith angle θ. It is expressed in radians.

Shower size S(1000). The signal shape exhibits also a dependence on the energy of the primary particle. For the SD events, the energy estimator is S(1000), the signal interpolated at 1000 m from the shower axis. The reconstructed energy E_{SD} is computed from S(1000) taking into account the attenuation in atmosphere depending on the zenith angle and using a parameterization calibrated with a subset of hybrid events, i.e. events detected also by the fluorescence telescopes. Since the relationship between S(1000) and E_{SD} is calibrated with data and photons are characterised by a different shower development and content, the estimated energy of photon-induced showers appears to be significantly overestimated. Therefore, here we choose not to use the reconstructed energy E_{SD}, but to use $\log_{10}(\mathrm{S}(1000)/\mathrm{VEM})$ as an input. Also in this case, the values are always of the order of unity.

Total number of triggered stations. All the considered simulated events have at least three triggered stations, but some of them exhibit up to tens of them. Some preliminary tests have already shown that including the information from more stations could possibly improve the discrimination power of the model. However, such an extension to higher numbers of considered stations would require a careful treatment of the missing information in the events with less triggered stations than the chosen considered number, which is beyond the scope of this baseline analysis. Therefore, we chose to postpone possible investigations to future works and to include here the total number of triggered stations N_{stat} as an input. The logarithm with base ten of N_{stat} is considered, to have values of the order of unity.

Simulated photon and proton events are labelled with $y_{\mathrm{true}} = 1$ and $y_{\mathrm{true}} = 0$, respectively. The number of input variables for each event is 463. Our goal is identifying the features and patterns that allow us to assign to events a predicted label y_{pred}, with values between zero and one, as close as possible to the true value y_{true}.

3. Convolutional Neural Network

3.1 Architecture

CNNs are well suited for structured data with symmetries. They search for the symmetries and properties in the data preserving the global structure. At each layer, the input is scanned with filters to create a feature map, which depends on the locations and weights of the detected features. A sketch of the network used for this analysis is shown in Fig. 2. The Exponential Linear Unit (ELU) [9] is used as the activation function in all the layers apart from the last one. Regularization techniques are applied throughout the network to reduce the risk of overfitting.

PoS(ICRC2023)191

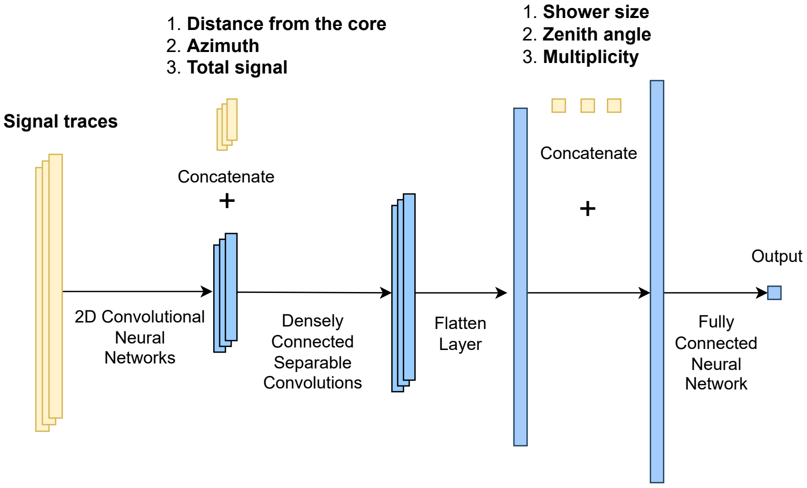

Figure 2: Sketch of the CNN. The input variables are added to the network at different stages by concatenating them to the previously obtained feature maps. The information at the level of individual stations is analyzed with convolutional layers, then fully connected layers are used to include the information at the event level.

In the first part of the network, the signal traces given by the three detectors registering the highest signals in each event are analyzed. Five 2D convolutional layers are applied without mixing the information from the individual stations and the resulting feature map is then concatenated with the additional input variables at the station level (distance form the shower core, azimuth in SP coordinates and total signal). As a second step, all the inputs at the station level are merged with densely-connected separable convolutional layers. In densely-connected networks the performances are improved by connecting the output of each layer to the input of all the subsequent ones, allowing to re-use and combine the features found at different depths [10]. The computational efficiency is increased by using separable convolutional layers, where the computation is split in two steps to reduce the number of parameters [11].

Finally, the last feature map is flattened and concatened with the information at the level of the event, to be analyzed with a fully connected neural network of three layers. The output node returns a value between zero and one by using a sigmoid activation function, which allows to classify events as photons (if the output is closer to one) or protons (if it is closer to zero).

3.2 Training strategy

We used the deep neural network library Keras [12], which runs on top of the TensorFlow [13] platform. The adaptive optmizer ADAM [14] with an initial learning rate of 10^{-3} is chosen and a binary cross entropy function is used for the binary classification. The training consists of 100 epochs, in which the training data set is split in batches of size 128. We use $\sim$ 73 000 events (60%) for the training and $\sim$ 24 000 events (20%) for the validation performed at the end of each epoch. After the training process, the final performance of the neural network is validated with a test set of $\sim$ 24 000 events (20%), which have not been used during the training. Each set consists of approximately 75% photons and 25% protons.[3]

[3] Since our data set is imbalanced, we have also verified if the results are modified by undersampling the majority class. No significant differences were found.

PoS(ICRC2023)191

A k-fold cross validation is exploited to better evaluate the performance of our model. The procedure consists of splitting the data set into k folds: $(k-1)$ folds are used to train (and validate) the model and the hold-out fold is used as the test set to evaluate the model; the procedure is repeated k times, so that each group is used once as the hold-out test set. However, our data set is imbalanced and randomly splitting our data could result in folds with few representatives from the minority class, which would produce predictions biased towards the majority class. Therefore, we use a stratified k-fold cross-validation, which ensures that the proportion of the representatives of the two classes found in the original distribution is respected in all the folds. In the end, a total of $k = 5$ models are fitted and evaluated each time on a different test set, and the overall performance of the network is calculated as the mean of the obtained evaluation scores.[4]

4. Results

The training process is summarised in Fig. 3, where the loss and the accuracy are shown as a function of the epoch for each one of the $k = 5$ models. The cross entropy loss function measures the difference between the values predicted by the model and the true ones, whereas the accuracy is defined as the percentage of correct predictions in the training set. Loss and accuracy are also computed on the validation sample at the end of each epoch, to evaluate the model performance during the training. For the sake of clarity, the validation metrics are shown only for one model: as expected, they oscillate around the corresponding metrics computed on the training set.

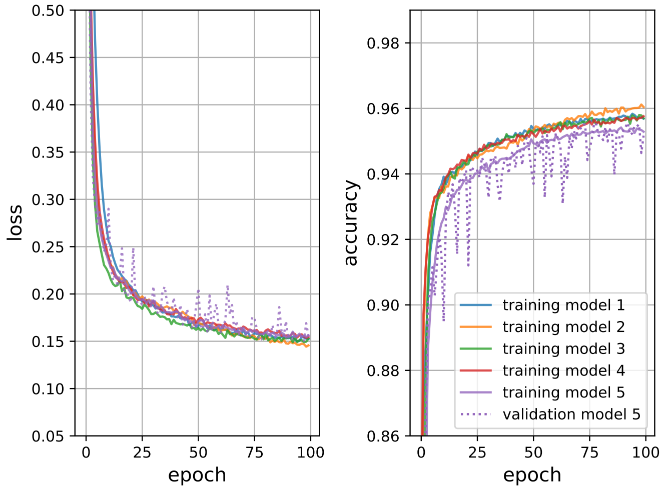

Figure 3: The loss and the accuracy per epoch computed on the training sample for each one of the models corresponding to the $k = 5$ groups (solid lines). For one of the models also the validation loss and accuracy are shown as an example (dotted line).

At the end of each training, the model is evaluated on the hold-out fold, to compute the corresponding scores. The overall results are summarized with the mean of the scores over the five trained models. The mean accuracy of our network is then $(95.64 \pm 0.39)\%$.

However, since we are dealing with an imbalanced data set, the accuracy could lead to overestimating our trained network performance and it is thus better to assess it by computing the balanced accuracy. It is defined as the arithmetic mean of sensitivity (the correctly identified positives over the total positives) and specificity (the correctly identified negatives over the total negatives). For our network we obtain $(94.62 \pm 0.24)\%$.

The default value for the decision threshold that is used to compute the accuracy is $y_{\text{pred}} = 0.5$. The distribution of predicted labels y_{pred} computed on the test sample is shown in the left plot of Fig. 4 for one of the five models, as an example. The decision threshold is represented by a vertical

[4] The choice of $k = 5$ implies the partition into training (60%), validation (20%) and test (20%) sets mentioned above.

PoS(ICRC2023)191

line. As expected from the high value of accuracy we obtain, the photon distribution and the proton one exhibit a peak around one and zero, respectively.

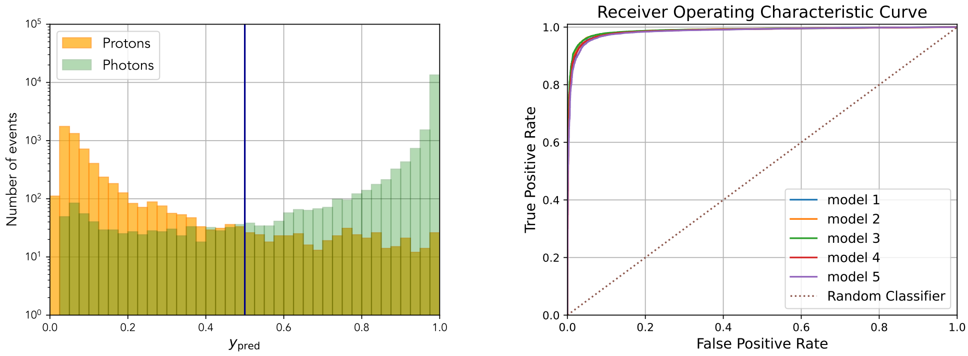

Figure 4: Left: The distribution of predicted labels y_{pred} computed on the test sample for one of the five trained models, as an example. Right: ROC curves obtained for the five models trained during the cross-validation procedure (see the text).

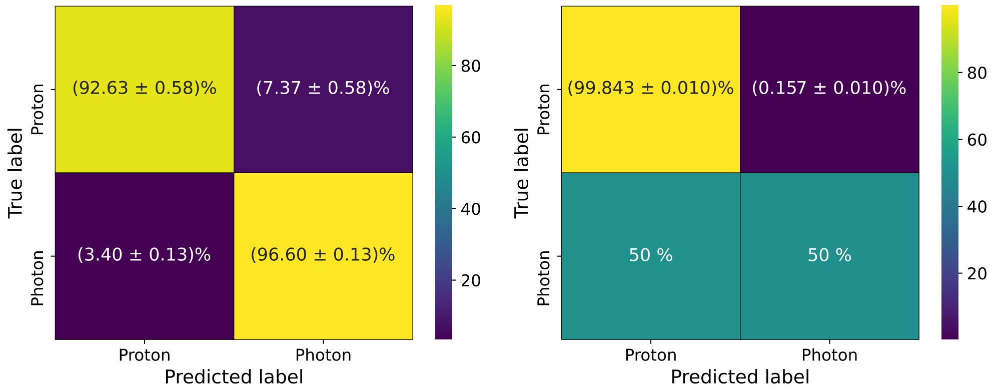

Figure 5: The confusion matrix obtained from the mean over the five models of the metrics obtained with the default decision threshold of $y_{\text{pred}} = 0.5$ (left) and with a decision threshold of $y_{\text{pred}} = 0.99$ (right); the uncertainty is given by the error on the mean.

A common way to summarize the performance of a classifier is that of using the confusion matrix. In Fig. 5 the tables diplay the mean percentages of proton and photon events in the test sample which have been correctly and wrongly classified. The confusion matrices are normalized over the rows, i.e. with respect to the true labels of the events; hence, each box shows the rate of true/false predictions in the predictions of a given class (the one of negatives, protons, or the one of positives, photons). Since the cross-validation we implemented provided five values for each metric, the mean values with their uncertainties are reported. On the left panel of Fig. 5 the metrics refer to a decision threshold of $y_{\text{pred}} = 0.5$, the one also used to compute the accuracy during the training process.

The metrics we are mainly interested in to assess the performance of a proton/photon classifier are the ones reported in the upper row: the true negative rate (TNR), shown on the left corner of the confusion matrix and also called background rejection, and the false positive rate (FPR), shown on the right and also referred to as background contamination. The signal efficiency is

PoS(ICRC2023)191

the true positive rate (TPR), that is the rate of photons which are correctly classified, shown on the lower right corner of the matrix. For comparison with other analysis, we want to evaluate the metrics at the decision threshold corresponding to a signal efficiency of 50%. The receiver operating characteristic (ROC) curve, like the ones in the right panel of Fig. 4, shows the signal efficiency against the background contamination at various decision threshold settings. The larger is the area below the ROC curve, the better is the performance of the classifier. From the ROC curves we computed that the decision threshold to have a TPR of 50% is $y_{\text{pred}} = 0.990 \pm 0.001$. For the confusion matrix in the right panel of Fig. 5 the decision threshold corresponding to a TPR of 50% is used ($y_{\text{pred}} = 0.99$); the background contamination and the background rejection are (0.157 ± 0.010)% and (99.843 ± 0.010)%, respectively. The achieved results are very close to the background rejections at 50% signal efficiency found in recent Auger analyses with hybrid events [2], which is about 99.9%.

5. Conclusions and Outlook

In this preliminary study, we developed a CNN algorithm that successfully classifies simulated photon/proton vertical (θ<60°) air-showers with energies above $10^{18.5}$ eV, by using the detector response of the Auger SD as a case study. The background rejection at 50% signal efficiency we achieve is (99.843 ± 0.010)%. The results are promising, especially considering that a classification power similar to the ones obtained with hybrid events is reached with simulated SD events alone, that is with less information available but potentially larger statistics at the highest energies when it is applied to real events. Further improvements of the performances are expected, e.g. by including more information from the SD array or by increasing the number of the simulated events. Besides, deep neural networks are also particularly suitable to combine data from different detectors, which would be the case when the simulated response of the AugerPrime [15] upgrade is included, possibly further improving the discrimination power with additional information on the events.

Acknowledgments

We wish to thank Jonas Glombitza for the useful inputs and fruitful discussion.
This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 945422.

References

[1] Pierre Auger Coll., *NIM A*, **798** (2015) 172.

[2] Pierre Auger Coll., *Universe, 8 (2022) 579*

[3] Pierre Auger Coll., *Phys. Rev.* **D 96** (2017) 122003.

[4] Pierre Auger Coll., *JCAP* **05** (2023) 021.

[5] D. Heck et al., **FZKA-6019** (1998) .

[6] T. Pierog et al., *Phys. Rev.* **C 92** (2015) 034906.

[7] A. Ferrari et al., Report CERN-2005-10.

[8] S. Argirò et al., *NIM A*, **580** (2007) 1485.

[9] D.A. Clevert et al., arXiv.1511.07289 (2015)

[10] G. Huang et al., arXiv:1608.06993v5 (2018).

[11] F. Chollet, arXiv:1610.02357v3 (2017).

[12] F. Chollet, Keras website.

[13] M. Abadi et al., TensorFlow website.

[14] D. P. Kingma and J. Ba, arXiv:1412.6980v9 (2017).

[15] A. Castellina for the Pierre Auger Coll. , *Web Conf.* **210** (2019) 06002.

PoS(ICRC2023)191