PROCEEDINGS OF SCIENCE

Machine learning-based analyses using surface detector data of the Pierre Auger Observatory

Steffen Hahn[a,*] **for the Pierre Auger Collaboration**[b]

[a] *Karlsruhe Institute of Technology - Institute for Astroparticle Physics, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany*

[b] *Full author list:* http://www.auger.org/archive/authors_2024_11.html

E-mail: spokesperson@auger.org

The Pierre Auger Observatory is the largest detector for the study of extensive air showers induced by ultra-high-energy cosmic rays (UHECRs). Its hybrid detector design allows the simultaneous observation of different parts of the shower evolution using various detection techniques. To accurately understand the physics behind the origin of UHECRs, it is essential to determine their mass composition. However, since UHECRs cannot be measured directly, estimating their masses is highly non-trivial. The most common approach is to analyze mass-sensitive observables, such as the number of secondary muons and the atmospheric depth of the shower maximum.

An intriguing part of the shower to estimate these observables is its footprint. The shower footprint is detected by ground-based detectors, such as the Water-Cherenkov detectors (WCDs) of the Surface Detector (SD) of the Observatory, which have an uptime of nearly 100%, resulting in a high number of observed events. However, the spatio-temporal information stored in the shower footprints is highly complex, making it very challenging to analyze the footprints using analytical and phenomenological methods. Therefore, the Pierre Auger Collaboration utilizes machine learning-based algorithms to complement classical methods in order to exploit the measured data with unprecedented precision. In this contribution, we highlight these machine learning-based analyses used to determine high-level shower observables that help to infer the mass of the primary particle, with a particular focus on analyses using the shower footprint detected by the WCDs and the Surface Scintillator Detectors (SSD) of the SD. We show that these novel methods show promising results on simulations and offer improved reconstruction performance when applied to measured data.

7th International Symposium on Ultra High Energy Cosmic Rays (UHECR2024)
17. – 21. November 2024
Malargüe, Mendoza, Argentina

∗Speaker

ISSN 1824-8039. Published by SISSA Medialab.

PoS(UHECR2024)091

1. Introduction

Cosmic rays above an energy of 1 EeV are commonly referred to as ultra-high-energy cosmic rays (UHECRs). Due to their low flux [1], the direct detection of UHECRs is not feasible. However, when UHECRs interact with the atmosphere, they produce a cascade of secondary particles commonly referred to as extensive air showers (EAS). The Pierre Auger Observatory is the largest detector for the study of extensive air showers induced by UHECRs. It is located in Argentina, in the Province of Mendoza, and is designed to simultaneously detect EAS using a hybrid technique. One of the main scientific goals of the Pierre Auger Observatory is to understand the physics behind the sources of UHECRs. Having orders of magnitude higher energy than any particle accelerated by human-made devices, the origin of UHECRs must be attributed to the most extreme processes in the Universe. To gain insight into these processes and to potentially identify them, an important piece of information is the mass composition of UHECRs. Analyses in this regime are based on so-called 'mass-sensitive observables' (MSOs). Essentially, MSOs are properties of EAS that statistically encode the mass of the primary particle. The most promising of these MSOs are the depth of the shower maximum X_{max} and the number of muons N_{μ} produced in the EAS [2]. Since the former is related to the electromagnetic and the latter to the hadronic part of the shower cascade both observables jointly contribute to the determination of the mass composition.

This contribution focuses on vertical showers with zenith angles $\theta \leq 60^{\circ}$ simulated for and observed by the Pierre Auger Observatory. In this zenith range, the direct detection of X_{max} and N_{μ} is based on the Fluorescence Detector [3] and the Underground Muon Detector [4], respectively. However, both of these specialized detector systems are restricted by uptime and coverage. Such a reduction in exposure can be mitigated by inferring shower observables from the measurements of the ground particle distributions by the surface detector (SD) stations of the Pierre Auger Observatory which have near-perfect uptime. The SD consists of various triangular grids of regularly spaced, autonomous detector stations that sample the distribution of particles of an EAS at ground level, referred to as shower footprint. Initially consisting only of water-Cherenkov detectors (WCDs), new additional particle detectors, such as the surface scintillator detector (SSD), have been added to the SD stations during a major upgrade of the Observatory named AugerPrime [5].

Exploiting the complex spatio-temporal information of the shower footprint using classical approaches is incredibly difficult and requires an accurate understanding of the universal behavior of air showers and the correct modeling of the spatio-temporal signal distribution at ground level [6]. Nevertheless, using data-driven approaches mitigates the need to find all the relevant information stored in the shower footprint. A simple way to work data-driven is by using neural networks (NNs) uncovering important features during the training process. Due to their modularity and adaptability NNs allow for an easy mapping of non-trivial shower footprint inputs to high-level shower observables.

1.1 Base neural network architecture

At the time of writing, all NNs used for the shower-by-shower reconstruction of high-level shower observables from SD data share the same 'architectural' ideas. Although, the elementary building blocks of all NNs differ, each NN is essentially composed of three separate sub-NNs that are optimized for addressing subtasks.

The first sub-NN (sNN1) essentially compresses the information of the time signals measured by the various detectors of the SD stations, called traces, into a small set of (scalar) trace features. Two stations triggered by the same EAS have a high probability of measuring different parts of the developing shower. There is little additional information from correlating signal bins of multiple traces, even if the traces are from neighboring stations. Hence, each trace, coming from the same detector, is – in general – treated and processed by sNN1 in exactly the same way. The architecture of sNN1 normally resembles NNs used for audio processing or language interpretation. The features computed from the traces are then added to a mixture of handcrafted features derived from reconstructed observables, e.g., the standardized relative trigger times. The resulting set of features is then[1] encoded in a 3D tensor where the first two dimensions correspond to the relative positions in the SD of the triggered stations. Normally, the station with the largest signal is placed in the center of the spatial coordinates of the 3D tensor. An example of an encoding procedure is outlined in [7]. The 3D tensor is then used as the input for the second sub-NN (sNN2). How the encoding is done and which additional features are selected depends on the analysis and the architecture of sNN2. The architecture of sNN2 is chosen in such a way that the sets of features of all triggered detectors can be correlated with each other. Thus, the architecture is designed in a similar way to that of NNs used in image recognition. In this analogy, the 'color channels' correspond to the found and selected features. The output of this sub-NN can the be used as input for one (or more) sub-NNs which combine all the information to predict one (or more) of the desired shower observables.

1.2 Simulated data sets

The training and test data sets used for each study presented in the following sections are not exactly the same. Although, the data sets differ in size, the preprocessing procedures, and the choice of inputs, there are commonalities. Namely, all data sets consist of detector simulations, performed with Offline [8], of air shower simulations, performed with CORSIKA [9]. The primary particle distribution follows an equal mixture of four primaries (p, He, O, Fe). The simulated showers are uniformly distributed in $\sin^2\theta$ ($\theta \in [0°, 65°]$) and in logarithmic energy ($\lg(E/\mathrm{eV}) \in [18.0, 20.2]$). For each study the training and test sets are non-intersecting. More detailed information of the used data sets can be found in the corresponding primary sources.

2. Neural networks for the reconstruction of shower observables

In general, the setup presented in Section 1.1 is applicable to any shower observable assuming that there is useful information encoded in the footprint of an EAS. However, to estimate the performance of NN-based estimators it is convenient to focus on NNs predicting shower observables which can be estimated using standard reconstructions methods or observed by specialized detectors.

2.1 Estimation of shower energy

The standard energy estimator of the SD-1500 depends only on a signal estimate at a predefined distance to the shower axis and the reconstructed zenith angle of the air shower. The free

[1] Normally, the traces are already encoded into a similar tensor simplifying the NN architecture. However, this is technically not necessary since the encoding only matters in the second part of the NNs.

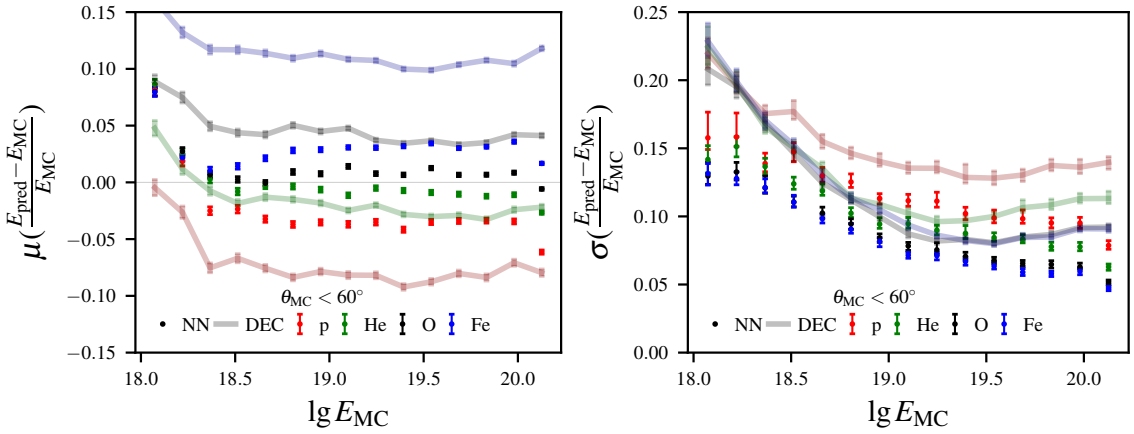

Figure 1: Bias (*left*) and resolution (*right*) of the relative energy error as a function of the logarithmic Monte-Carlo energy for different primary particles for the DEC (see Section 2.1) and the NN-based energy predictor. The underlying data are uniformly distributed in logarithmic energy and use equal proportions for the different primary particles. The NN-based predictor has a smaller inter-primary bias and better resolution than its classical counterpart. Kindly provided by Fiona Ellwanger.

parameters of the functional form of the estimator are gained from likelihood minimization using showers detected by the FD and SD simultaneously, called hybrid events, making it independent of air shower simulations. Details can be inferred from [10]. Hence, to make a fair comparison between an NN-based and the standard energy estimator on simulations, it is mandatory to create a similar energy estimator on simulations. Such an estimator can be determined from a (direct) global fit to simulated EAS using the functional form of the standard energy estimator. This new baseline model [11] is denoted as direct energy calibration (DEC).

Due to the mass-energy degeneracy and the simplicity of the method, the DEC exhibits a strong primary-dependent bias of up to ±10%, assigning on average a lower energy to protons and a higher energy to iron (see Fig. 1). By using an NN-based approach this bias (at least) is halved. At the same time the standard deviation of the relative energy error, a measure for accuracy, is improved by more than 3 percentage points over the whole energy range. Thus, the NN prediction accounts better for the primary mass information contained in the shower footprint. Since an unbiased and more accurate energy estimate is crucial for the correct interpretation of MSOs, the NN-based model is preferred when considering only simulations.

However, due to shortcomings in both detector and shower simulations, e.g., [12], a direct application of the NN-based estimator on measurements is not straightforward. Even after calibration with direct measurements of the FD energy, the performance of the NN-based energy estimator decreases significantly. Using hybrid events, the energy predictions of the NN reproduces the FD energy measurements on a similar level as the standard energy estimator [11].

At the time of writing, it is still unclear how to adequately account for all systematic effects and differences between simulations and measurements in a general way. Finding procedures that allow to tap into the full potential of NN-based energy estimators seen in simulations is the subject of future studies. Nevertheless, [11] demonstrates that one of the main problems, when using NN-based approaches for data of the Pierre Auger Observatory, is the transition to real measurements

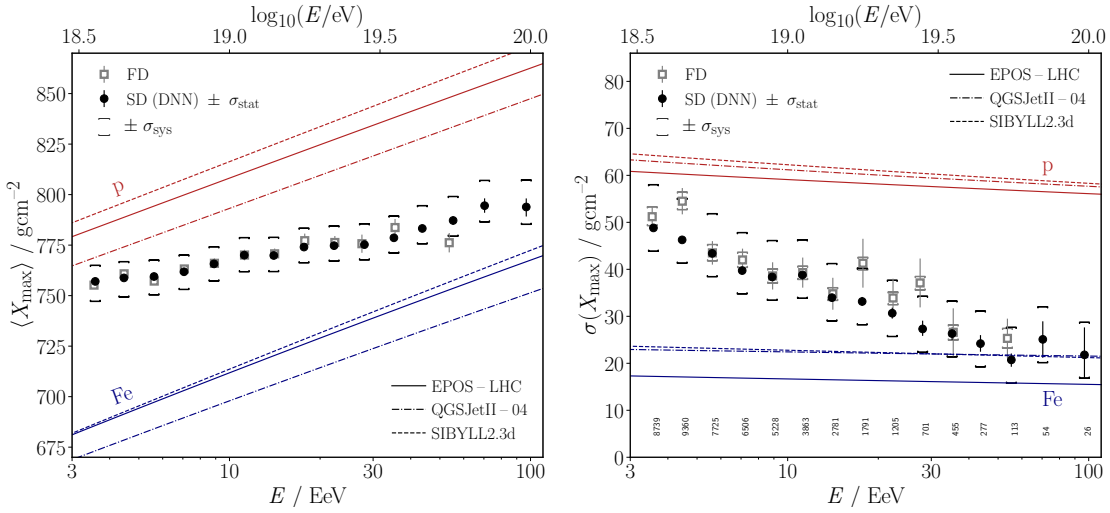

Figure 2: First (*left*) and second (*right*) moments of X_{max} as a function of reconstructed SD energy comparing the direct measurement (FD) and NN-based reconstruction [17]. The red and blue lines (of different line style) represent the predictions of the most commonly used hadronic interaction models (in order of appearance [14–16]) for pure proton and iron compositions, respectively. The reconstructed moments of both methods are in agreement with each other.

which has to be handled carefully.

2.2 Estimation of depth of the shower maximum

The NN-based approach in Section 2.1 demonstrates that NNs exploit the mass information contained in the footprint improving the primary-dependent bias. Hence, it is reasonable to develop NN-based estimators that target MSOs directly, mitigating the need for complex, spatio-temporal modeling of ground signals [6]. The best shower observable to test this is the depth of the shower maximum X_{max}, since it is measured directly by the FD.

Using a much more refined architecture as in Section 2.1, [13] summarizes the performance of such an NN on simulations and on measurements. Again, due to the shortcomings in simulations, the NN-based approach requires corrections[2] to account for effects, such as yearly variation of atmospheric conditions, before being applied to measurement data. This removes unphysical modulations in the predictions. After all corrections are applied, the predictions of the NN are still shifted by about $30\,\mathrm{g\,cm^{-2}}$ when comparing to direct measurements of the FD. This shift can be attributed to the muon puzzle [12]. Since the shift is nearly constant over the whole energy range, it is calibrated away using hybrid events. The X_{max} predictions of the NN agree well with the X_{max} measurements of the FD in the first and second moments of the X_{max} distributions (see Fig. 2). Both methods show a transition from a lighter to a heavier composition of UHECRs.

Although, the systematic uncertainty is higher for the SD predictions, the main advantage of the SD-based estimation of X_{max} is the significant increase in statistics by a factor of about eight. The use of the NN predictions reveals additional breaks in the first moment that roughly correspond

[2]Note that all corrections are done in such a way to preserve the mean value of the predictions.

to established features in the UHECR spectrum [17]. Moreover, the additional data allows to extend the first and second moments beyond 50 EeV.

2.3 Impact of detector upgrade

In addition to well-performing reconstruction tools, machine learning-based predictors can also be used to study how the amount of information affects the reconstruction quality of shower variables [18]. Fig. 3 shows two separate comparisons of five independent machine learning-based estimators to reconstruct the relative[3] muon number R_μ, where the first entry, boosted decision tree (BDT), is the only method not based on NNs. On the left side of Fig. 3 is the linear correlation coefficient of the prediction and the Monte-Carlo value and on the right side the classification accuracy for all predictors separately. The convolutional neural network (CNN) in the middle acts as the baseline. The CNN uses encoded footprints as inputs and a convolution-based architecture for sNN1 (see Section 1.1). Thus, the spatial information of the footprint and the temporal information of the traces are accessible. The alternative NN (ALT) has a similar architecture like the CNN. However, the sNN1 is replaced by two commonly used trace features, the total signal and the rise time[4]. The inputs used for the BDT are simplified even more. They are only scalar footprint information concatenated to a one-dimensional vector. Both ALT and BDT perform worse than the CNN. Therefore, the spatial information in the shower footprint as well as the temporal information extracted directly from the features are taking into account in the CNN. In the recurrent neural network (RNN) sNN1 is replaced by stacked LSTM layers allowing a more refined trace feature extraction. Again, this impacts the overall quality of R_μ predictions. Due to the flexibility of the NNs, new measurement devices on the SD detector stations, such as the scintillator detectors, can be easily added to the reconstruction pipeline in the same way as the traces of the WCDs. Supplementing the RNN additionally with data from SSDs (see Fig. 3, marked with Phase-II) the prediction quality of R_μ is improved even further. Therefore, on simulations the NN-based approaches benefit from the access to structured information, as well as from more sophisticated sub-NN architectures without having to fundamentally restructure the architecture used.

Using a different, more sophisticated architecture consisting of transformers [20] which predicts R_μ and $X_{\max}$ simultaneously, the improvement can be quantified by comparing the standard deviations of the residuals [19] which corresponds to a measure for the resolution. For the muon number R_μ the resolution of the predictions improves significantly over the whole energy range (see Fig. 4) for all primaries independently. Both the SSD and the WCD are particle detectors with different responses to different particles of the shower cascade. The improved resolution indicates that the simultaneous measurement of the SSD and WCD helps to disentangle the electromagnetic and muonic components of air showers. This conclusion is supported since the prediction quality especially improves for low zenith angles [19] where the planar design of the SSD reaches full efficiency. Adding another detector has – most likely – less impact on the $X_{\max}$ prediction, since $X_{\max}$ is a feature of the electromagnetic shower cascade that is encoded in the shower front and trigger timing. Still, the resolution slightly improves overall for the test data set driven by the resolution of He, O, and Fe.

[3]The relative muon number R_μ is computed by dividing the muon number N_μ with the number of muons expected from protons at the same zenith angle and shower energy.

[4]Time until a certain signal threshold is deposited.

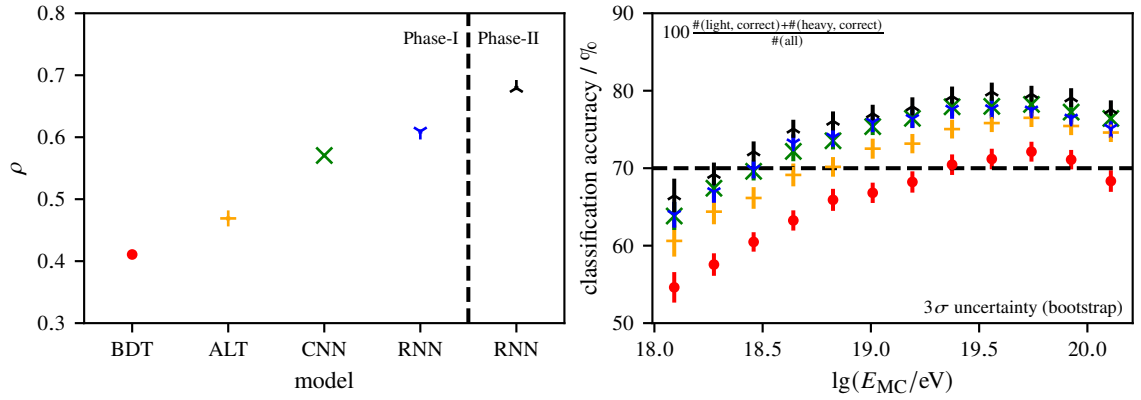

Figure 3: Linear correlation coefficient of NN predictions and MC value of R_μ (*left*) for different machine learning-based methods (see Section 2.3) and classification accuracy for the same methods as function of logarithmic Monte-Carlo energy. The more information of the footprint is used, the better the correlation and classification becomes. Taken from [18].

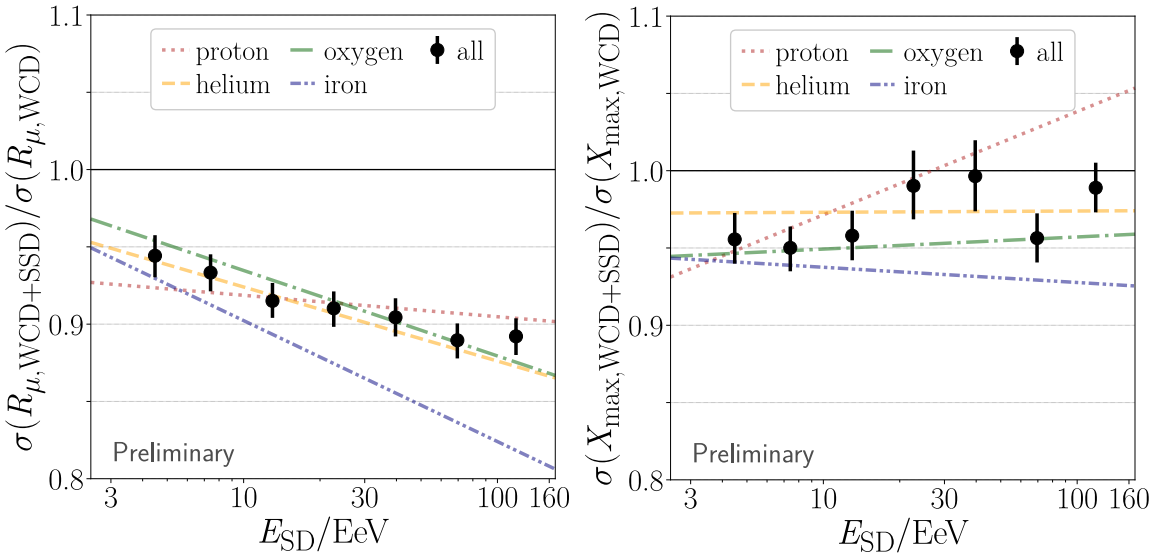

Figure 4: Fraction of standard deviation for the difference of R_μ (*left*) and $X_{\max}$ (*right*) predictions and Monte-Carlo values as a function of energy for NN-based approaches using the combined information of the WCD and SSD and using only the information of WCD. The straight lines are linear fits to the fractions of using only the pure compositions of the different primaries. Taken from [19].

3. Conclusion

In this contribution, the current state of machine learning-based analyses at the Pierre Auger Observatory was summarized, focusing on NNs that use SD information as inputs. It was shown that neural networks are powerful tools to extract high-level shower observables and that NN can even be used to probe the potential of newly deployed detectors, such as the SSD.

Although, the impact of systematic differences between measurements and simulations on NN-based predictions is not yet fully understood, NN-based analyses show exceptional performance on

simulations and promising results on measurements. Therefore, NN-based analyses are gradually being integrated into the data analysis pipeline of the Pierre Auger Observatory supplementing standard approaches and potentially replacing some parts of them in the near future.

References

[1] Aab, A. *et al.*, Phys. Rev. D **102**, 062005 (2020)

[2] Matthews J., Astropart. Phys. **22**, 387-397 (2005)

[3] Abraham, J. *et al.*, Nucl. Instrum. Meth. A **620**, 227-251 (2010)

[4] Aab, A. *et al.*, JINST **16**, P04003 (2021)

[5] Castellina, A *et al.*, EPJ Web of Conferences **210**, 06002 (2019)

[6] Stadelmaier, M. *et al.*, Phys. Rev. D **110**, 023030 (2024)

[7] Hahn, S. *et al.*, PoS **ICRC2021**, 239 (2021)

[8] Agiro, S. *et al.*, Nucl. Instrum. Meth. A **580**, 1485-1496 (2007)

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

[10] Aab, A. *et al.*, Phys. Rev. D **102**, 062005 (2020)

[11] Ellwanger F. *et al.*, PoS **ICRC2023**, 275 (2023)

[12] Albrecht, J. *et al.*, Astrophys. Space Sci. **367**, 27 (2022)

[13] Abdul Halim, A. *et al.*, ArXiv 2406.06319 (2024)

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

[15] Ostapchenko, S., Nucl. Phys. B Proc. **151**, 143 (2006)

[16] Riehn, F. *et al.*, Phys. Rev. D **102**, 063002 (2020)

[17] Abdul Halim, A. *et al.*, ArXiv 2406.06315 (2024)

[18] Hahn, S. *et al.*, PoS **ICRC2023**, 318 (2023)

[19] Langer, N. *et al.*, PoS **ICRC2023**, 371 (2023)

[20] Vaswani, A. *et al.*, NIPS **30**, 5999 (2017)