LIP-STUDENTS-22-01

Probing the cosmic ray composition with SWGO

Afonso Guerreiro[1,a]

[1] *Instituto Superior Técnico, Lisboa, Portugal*

Project supervisors: R. Conceição, B.S. González

October 3, 2022

Abstract. An investigation on the ability to identify the cosmic ray mass composition using SWGO simulation data combined with machine learning algorithms is made. The simulations of extensive air showers were performed resorting to CORSIKA. The energies of the resultant particles in the stations were converted into signals, normalised, and then fed to a convolutional neural network (CNN). The accuracy of the CNN is presented, for different normalisations, as a function of the Fill Factor (FF). The relation between muon numbers and proton probability is discussed as well. The results show that CNNs can be used to distinguish proton-induced showers from iron-induced showers, with reasonably good discrimination for values of fill factors as low as 1%. Furthermore, the classification process doesn't take into consideration the muon number and seems to be based on the shower footprint, solely.

Keywords: Extensive Air Showers, Mass composition determination, Experiment Fill Factor, Convolutional Neural Network

1 Introduction

1.1 Cosmic Rays

Cosmic Rays are energetic particles that hit the Earth's atmosphere at a rate of about 1000 per square meter per second [1]. They are mostly made up of hydrogen nuclei (around 90%), alpha particles (roughly 9%) and heavier nuclei. Their energies range from a few hundred MeV to 300 EeV [2]. Cosmic rays with energies ranging from around 10 GeV up to 100 PeV are expected to be produced in our galaxy. Those that can be attributed in their origin to the sun, have a strong temporal association with peaks in solar activity. The fact that the most energetic particles have gyro-radii of the size of the galaxy, seems to point to an extra-galactic origin. The mechanisms responsible for the acceleration of these particles are still not fully comprehended, the leading candidate being supernova explosions. However, for some of the most energetic particles, this hypothesis seems not to suffice, possibly opening the door to new astrophysical events.

1.2 Extensive Air Showers

When a primary particle, either a nucleus or a photon, reaches the upper atmosphere, a sequence of violent collisions is set into motion forming a cascade of particles called an extensive air shower. Such events can emit Cherenkov Light, as the particles that form the shower travel faster than the speed of light divided by the refractive index of air, as well as fluorescent light form the excitation of air molecules. On the other hand, if the energy of the primary particle or the altitude of the detectors is sufficiently high, then the secondary particles may be deleted at the ground, as is illustrated in Fig. 1.

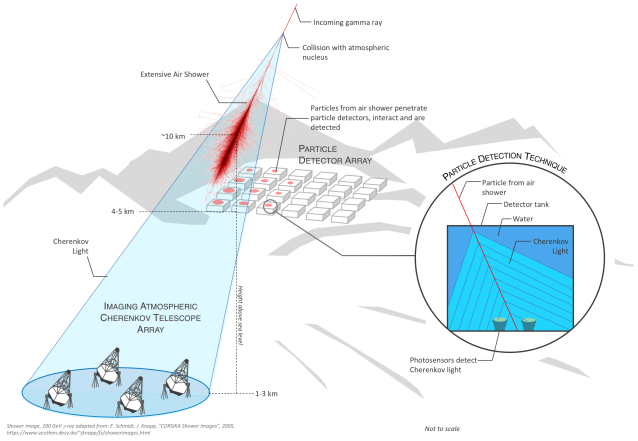

Figure 1. Depiction of an air shower and its detection resorting to Cherenkov Light and an array of particle detectors.

1.3 The SWGO

The Southern Wide-field Gamma-ray Observatory (SWGO) is a proposed experiment, located in South America at a latitude between 10 and 30 degrees South, at an altitude higher than 4.4km. No such instrument exists in the southern hemisphere, where the great potential exists for the mapping of large-scale emissions as well as providing access to the full sky for transient and variable multi-wavelength and multi-messenger phenomena [3]. Its main purpose is to study gamma rays with energies ranging from the hundreds of GeV to the tens of PeV. The observatory will be formed by several Cherenkov detector units with a high Fill-Factor (ratio of the area occupied by detectors to the total area) core detector, and a low-density outer array, as depicted in Fig. 2. When a particle crosses the detector, which is itself filled with water, it will radiate, via Cherenkov radiation. The resultant Cherenkov photons can then be picked up by sensitive photo-multipliers tubes (PMTs) placed inside

[a] e-mail: afonsojguerreiro@tecnico.ulisboa.pt

the detector unit, registering the passage of a particle. As the SWGO will focus on gamma rays, the cosmic rays constitute a background for this experiment. Nonetheless, the cosmic rays' mass composition is of great interest and as such, in this paper, the use of an algorithm to distinguish proton-induced from iron-induced showers is proposed.

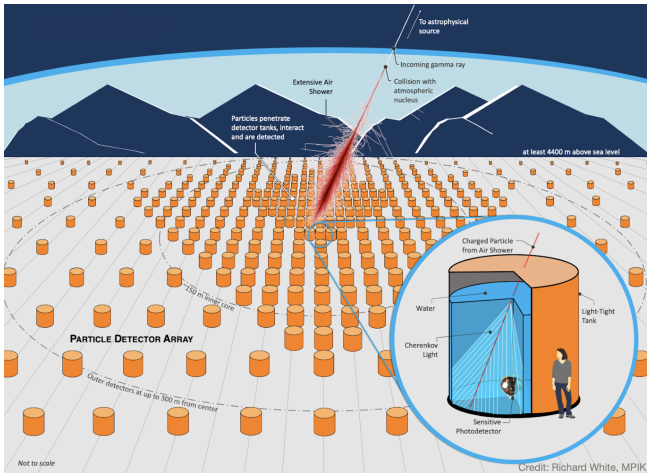

Figure 2. Schematics depicting the future SWGO and its functioning

2 Convolutional Neural Networks

A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. In recent years, due to their excellent performance and relevant applications in different fields, CNNs have become state-of-the-art in image recognition and signal analysis. They are mainly constituted by three different types of layers: Convolution, Pooling and Fully Connected Layers, as exemplified in Fig. 3.

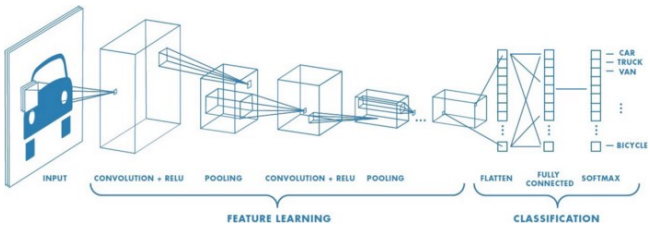

Figure 3. Schematic representation of a Neural Network

Convolution Layer

A convolution layer is a fundamental component of the CNN architecture that performs feature extraction [4] and is usually the first component of a neural network. In this layer, a small matrix called kernel is applied across an input tensor. In each iteration, the respective section and the kernel are multiplied using the dot product rule, being the output fed into an array. After each iteration, the region where the kernel acts is shifted by an amount known as a stride. The values of the kernel itself are learnable parameters that are set during training, however, the size and stride of the kernel must be fixed beforehand. One of the advantages of this process is reducing the number of learnable weights when compared to fully connected neural networks, thus increasing their efficiency.

Pooling Layer

The goal in this layer is to subsample the input image, i.e. reduce its size, in order to improve computational performance [5]. The most common way of doing so, and the one used in this paper, is called Max Pooling. Simply put, this method extracts the maximum value in each region and discards all the others, outputting this value onto an array. This process is repeated until the entirety of the image has been covered. It is worth mentioning, that there are no learnable weights nor biases in this stage.

Fully connected Layer

The resultant array is at this point flattened, meaning transformed into a 1-D array i.e. a vector, to be received as input by the fully connected layer. Here as the name suggests each node in the output layer connects with a node in a previous layer. Each neuron in this layer has an activation function, which determines if said neuron should be activated, as well as a weight, the latter being a learnable parameter determining the importance of that connection to the final result.

The structure used in the current paper is presented in Fig. 4. It consists of a convolution layer followed by a pooling layer, followed by the same pattern before flattening the signal and feeding into a series of fully connected layers.

Loss Functions

Loss Functions are critical for the good performance of an Artificial Neural Network (ANN), as it quantifies the deviation between a model's predictions from the correct results. For the purposes of this paper, the binary cross-entropy function was chosen, as it is most suited for binary classification problems, as is the case.

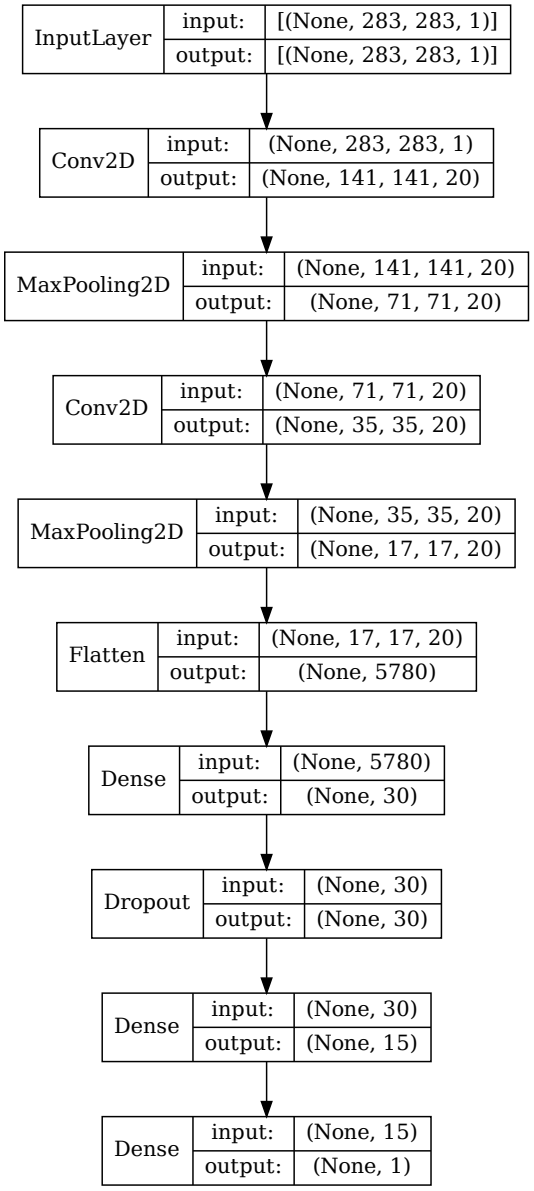

Figure 4. Structure of the Neural Network used in this paper

Data

Typically for machine learning purposes, the data is divided into three distinct categories, the first being the training data set, which is used to adjust learnable parameters as well as calculate the values of the loss function. The validation data is used to adjust hyper-parameters and evaluate the model during training, selecting the best model. Finally, the test data set is never used during training but instead to test the model's final accuracy and performance, with data not yet seen by the network. This division is summarised in Fig. 5

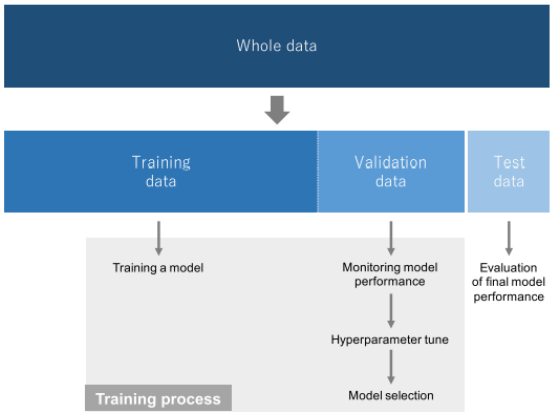

Figure 5. Schematic representation of the data organisation

3 Experimental procedure

The extensive air showers simulation was done by resorting to CORSIKA [6] software. The proton-induced showers were taken from a primary energy bin between 40 and 60 TeV, while iron was selected using a wider energy bin 10 to 100 TeV. Particle energies were collected in the cells and converted into signals using a parameterization obtained using a Geant4 [7] detector simulation. A cut on the total signal at the ground was imposed to emulate the shower energy reconstruction algorithm and have a fair comparison between the footprints at the ground level, as shown in Fig. 6.

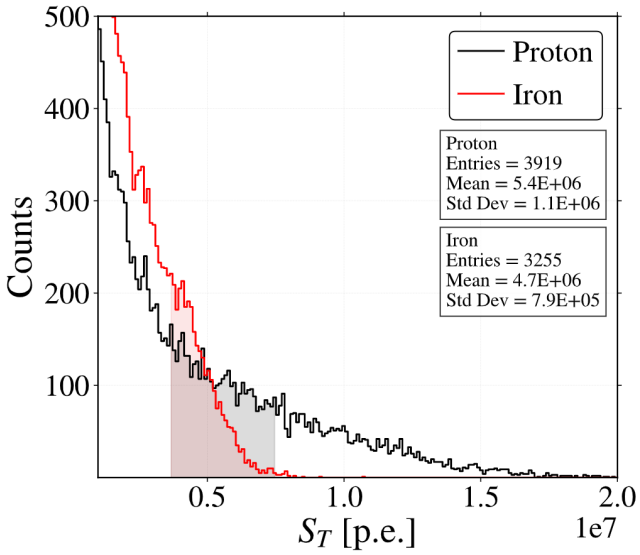

Figure 6. Counts as a function of the total signal at the ground. The shaded area depicts the events selected from the total simulated. The information in the boxes concerns the selected events only

The energies of the resultant particles in the stations were converted into a signal, producing 256x256 arrays where each pixel has associated with it the signal detected

in that station, as shown in Fig. 7 and Fig. 8. The station used for the energy to signal calibration is a water Cherenkov detector (WCD) with a radius of 2m, a water height of 1.7m, and instrumented with three 8" PMTs at the bottom of the station. This station is the current proposal of LIP for the future SWGO and it is known as Mercedes WCD [8].

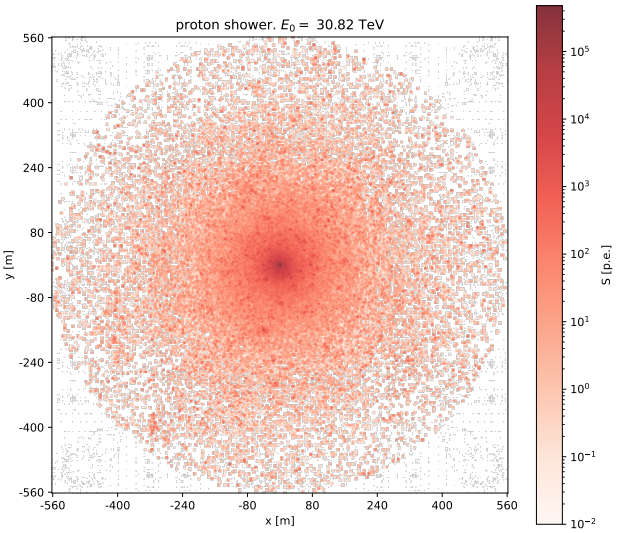

Figure 7. Proton induced shower footprint for a fill factor of 100%

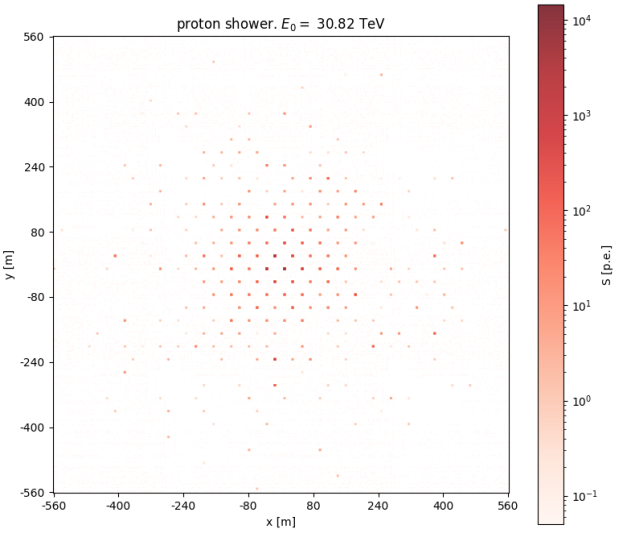

Figure 8. Proton induced shower footprint for a fill factor of 1%

To improve the generalisation capability of the network, two normalisation strategies were used. The first consisted in taking the natural logarithm of the signal and then bounding it to an interval from 0 to 10, which form henceforward will be referred to as the logarithmic normalisation. The binary normalisation consisted of setting all non-zero values equal to one. Since the latter has no information on the stations' signal intensity, the comparison between both normalisations provides some insight into the classification process, gauging if the results were based solely on the shape of the pattern produced on the ground or if the network was picking up information related to the event's calorimetric energy at the ground. The data set was then split in the manner presented in the previous section, such that the training data set encompasses 60% of the total, the validation data 15% and the test data 25%.The CNN was implemented resorting to the Keras [9] package for Python, the optimiser chosen was Adam, the batch size was 128 and the number of epochs was 30. The output of the CNN consisted of a value, ranging from 0 to 1, corresponding to the probability that the extensive air shower was initiated by a proton.

4 Results and Discussion

We begin by analysing the accuracy of the neural network, meaning, the percentage of times the network guessed the type of shower correctly as a function of the Fill Factor. We clearly expect that as the Fill Factor decreases so does the accuracy, as the signal gets fainter and fainter, the neural network should have more difficulty telling them apart. For that, the CNN was trained 10 times and the average value of the accuracy was taken which was plotted against the Fill Factor resulting in Fig. 9. The error associated with each data point consists of the standard deviation of the sample. The fluctuations from this trend can be attributed to the intrinsic stochastic nature of the process, as the minimisation process during training is only guaranteed to find a local minimum, meaning that fluctuations are to be expected. Given that the accuracy of the binary normalisation is in general lower than the one obtained with the logarithmic normalisation it becomes clear that information both on the morphology of the event and the station's signal is used in the classification process. Moreover, the accuracy remains relatively high down to fill factor values of 1%, and even for fill factors of 0.1% the CNN is right 3 out of 4 times

The number of muons is very sensitive to the composition of the shower, such that for iron showers it is expected to be larger when compared to proton-induced showers. However experimentally the measurement of the number of muons is quite complicated, making the detectors expensive or the muon tagging resolution bad. Nonetheless, as we have access to the muon number data we can use it as a consistency check and evaluate how it relates to the proton probability, as was done in Fig. 10. We observe two disjoint sets of points with no visible correlation or trend between them. This seems to imply that the CNN is not performing the discrimination using features related to the number of muons at the ground, otherwise we would expect a clear relation between muon number and proton probability.

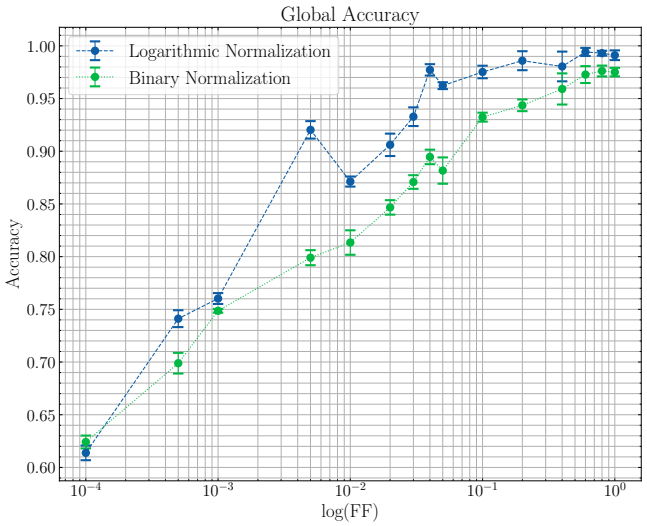

Figure 9. Average accuracy as a function of the logarithm of the Fill Factor

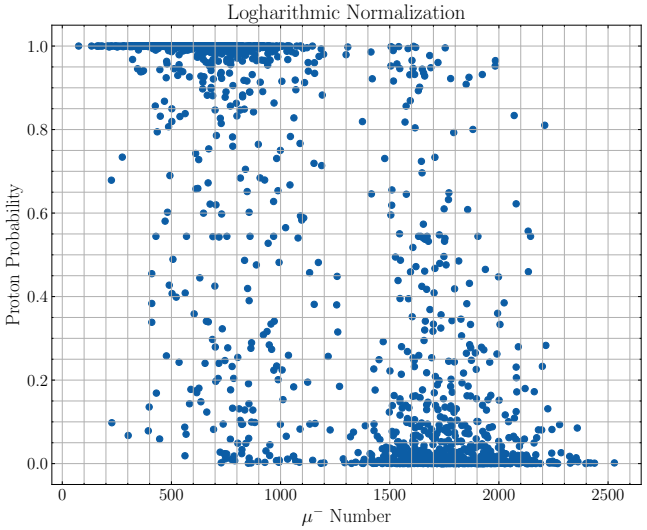

Figure 10. Proton Probability as a function of the muon number for FF=0.1%

5 Conclusion

In conclusion, we have shown that machine learning algorithms, in particular, CNNs can be used to classify proton and iron showers and that good discrimination can be attained with fill factors as low as 1%. Furthermore, it would appear that the CNN is oblivious to the muon content of the shower and is instead basing its classification on the shower footprint, meaning the classification is drawing information both from the shape produced on the ground as well as the station's signals, but not the muon content of the signal.

Acknowledgements

I deeply appreciate the opportunity to participate in the LIP summer internships program, as well as all the time and attention devoted by the LIP community and, in particular, Prof. Rúben Conceição and Prof. Borja Serrano González, whose help and dedication were essential to this work.

References

[1] T.K. Gaisser, R. Engel, E. Resconi, *Cosmic rays and particle physics* (Cambridge University Press, 2016)

[2] P. Biermann, G. Sigl, Lect. Notes Phys. **576**, 1 (2001), astro-ph/0202425

[3] *Swgo*, https://www.swgo.org/SWGOWiki/doku.php

[4] R. Yamashita, M. Nishio, R. Do, K. Togashi, Insights into Imaging **9** (2018)

[5] A. GERON, *Hands-on machine learning with scikit-learn, Keras, and tensorflow: Concepts, tools and techniques to build Intelligent Systems* (O'Reilly, 2019)

[6] D. Heck, *Corsika: A Monte Carlo code to simulate extensive Air Showers* (Forschungszentrum, 1998)

[7] S. Agostinelli et al. (GEANT4), Nucl. Instrum. Meth. A **506**, 250 (2003)

[8] P. Assis, A. Bakalová, U.B. de Almeida, P. Brogueira, R. Conceição, A. De Angelis, L. Gibilisco, B.S. González, A. Guillén, G. La Mura et al., *The mercedes water cherenkov detector* (2022), https://arxiv.org/abs/2203.08782

[9] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin et al., *TensorFlow: Large-scale machine learning on heterogeneous systems* (2015), software available from tensorflow.org, https://www.tensorflow.org/