ISTITUTO NAZIONALE DI FISICA NUCLEARE

Sezione di Catania

INFN/AE-07/01
23 Gennaio 2007

A NEURAL NETWORK APPROACH TO HIGH ENERGY COSMIC RAYS MASS IDENTIFICATION

Simone Riggi, Rossella Caruso, Antonio Insolia, Mario Scuderi
Dipartimento di Fisica e Astronomia, Università di Catania and INFN, Sezione di Catania

Abstract

An event-by-event study, based on neural network methods, of the mass identification in high energy cosmic rays was carried out with simulated data, in order to check the possibility of analyzing real data measured at the Pierre Auger Observatory. Extensive air showers were simulated with the CORSIKA code, using the hadronic model QGSJET98. The goodness of the method in recognizing the mass of the primary was tested making use of the parameters extracted from the simulated longitudinal profiles. We showed that the designed supervised neural network is able to discriminate, with high identification efficiency and purity, between proton- and iron-induced showers. We tested our method also in presence of a four components primary flux (proton, helium, oxygen, and iron). Typical results for the classification matrix obtained are presented and discussed.

PACS:96.50.sb,96.50.sd,84.35.+i

Published by **SIS-Pubblicazioni**
Laboratori Nazionali di Frascati

1 Introduction

Mass composition analysis is a fundamental task to test any theoretical model concerning the origin and the nature of the primary cosmic ray radiation at the highest energies. Different energy spectra are predicted to be observed at ground by the present theories, according to the mass of the primary particle, so the knowledge of the energy spectra for every mass component, or at least for groups of components, is required in order to discriminate among the proposed models.

At lower energies ($0.1 \div 100$ TeV) the composition of cosmic rays can be measured using direct detection techniques, such as spectrometers and calorimeters: the experimental data in this energy range tell us that the radiation is approximately made up of 50% protons, 25% α particles, 13% CNO and iron nuclei [1].

At higher energies, the measurement of the mass is generally performed by indirect techniques, which make use of parameters sensitive to the primary mass, and determined by the shower development in the atmosphere. Among such parameters, X_{max} (the depth at which the longitudinal shower has its maximum), N_{max} (the number of shower particles at X_{max}) and N_{muons} (the number of muons at a given distance from the shower axis) are widely used. In the knee region ($10^{15} \div 10^{17}$ eV) a recent analysis from KASCADE experiment, based on the deconvolution of a 5-component mass spectra starting from the experimental N_{max}-X_{max} scatter plots, shows that the knee is due to a decrease of the light component with respect to the heavier one, and that the knee position for higher masses shifts towards higher energy [2]. A clear increase of the mean logarithmic mass as a function of the primary energy is found in other experiments, such as EASTOP-MACRO [3]. While the experimental results are clearer in this intermediate energy region, the situation becomes controversial moving to the highest energies ($> 10^{17}$ eV): the HiRes analysis [4], based on the elongation rate method, the Yakutsk analysis [5], based on the comparison of experimental X_{max} distributions to QGSJET simulated ones, and the AGASA analysis [6], based on the comparison of experimental muon number distributions with simulated ones, suggest a composition dominated by the proton component. Recent re-analyzed data from Volcano Ranch [7] and Haverah Park [8] experiments, based on the comparison of the steepness parameter distributions, extracted from the lateral distribution function, with simulated ones, claim for a composition dominated by the iron component.

Interesting attempts to compare these results are found in [9] and [10]. Measurements from different experiments are difficult to compare, because the predictions are strongly dependent upon the hadronic models used in the analysis. These controversial results suggest that the problem of mass composition at the highest energies is still open and debated.

This is indeed one of the main objectives of the Pierre Auger Experiment, which consists of two observatories of about 3000 km^2 each, located at sites in the Southern and Northern hemispheres. The Southern Observatory is actually expected to be completed in a few months and is taking data as the deployment goes on. Two systems of detectors have been mounted to measure the shower properties: the surface detector (SD) consists of a grid of Cherenkov water detectors, measuring the particle density at ground level, hence the lateral distribution of the shower; while the fluorescence detector (FD) measures the fluorescence light emitted by the shower particles traversing the atmosphere, and the longitudinal profile of the shower [11].
Two kinds of approaches can be used to perform a composition analysis: the event-by-event approach uses pattern recognition methods, working with a set of shower parameters sensitive to the mass, in order to estimate the probability of identifying the mass of every observed event; methods of unfolding or deconvolution allow to infer the energy spectra for different mass components, starting from a data set of shower parameters, without any care regarding the mass of the single event.
It is clear that a mass identification study must be necessarily restricted to limited mass groups, since the absence of features strongly correlated with the primary mass and the presence of stochastic shower-to-shower fluctuations in the shower parameters, make a complete analysis very inefficient. The first approach could become inadequate, even with a powerful pattern recognition method, especially with a too large number of mass components. Keeping in mind these difficulties, an event-by-event reconstruction is anyway necessary if one wants to study possible correlations with other analysis, e.g. if one wants to correlate the mass of an event with its astrophysical arrival direction.
For these reasons we present in this paper the results of an event-by-event study, performed with parameters extracted from simulated showers and measurable with the FD detector at the Auger Observatory and with a neural network as identification tool.
The paper is organized as follows: section II describes the data set, built from CORSIKA simulations of extensive air showers, and the search for parameters sensitive to the mass. Section III presents the neural network which was designed and its application to simulated data. Section IV, finally, shows the obtained results and our conclusions.

2 The simulated data

The present study is based on a sample of simulated showers, which were generated with CORSIKA 6.002 [12], using QGSJET98 [13] as hadronic interaction model. Simulations were performed at the Lyon Computer Centre.
The CORSIKA system is one of the widely used codes for EAS simulation currently in

use. All the relevant particles and interactions are taken into account during the simulations, and a number of observables are recorded; among them, the longitudinal and lateral profiles of the showers, the arrival time distributions, and detailed lists of particles reaching the ground level.
Two kinds of simulated data set were used, with the following features:

- *Set I*: 4578 proton and iron showers, generated with an optimum 10^{-5} thinning, a power-law energy spectrum ($\gamma = 2$) in the range $10^{18} - 10^{20}$ eV, zenith angles generated according to the distribution $dN \propto \sin\theta\cos\theta d\theta$ in the range $0^{\circ} - 60^{\circ}$ degrees;

- *Set II*: 500 proton, helium, oxygen and iron showers, generated with an optimum 10^{-6} thinning, at fixed primary energies 10^{18}, $10^{18.5}$, 10^{19}, $10^{19.5}$, 10^{20} eV (100 events for each energy) and zenith angle fixed to 0°. Similar sets are available for 18°, 26°, 37°, 45° and 60°. For proton and iron events, a 53° zenith angle set was also available.

We made use of the amount of information contained in the simulated longitudinal curves, sampled in 5 g/cm^2 bins by CORSIKA, with the only request of limiting the profiles in the range 200-870 g/cm^2, these upper and lower limits being determined respectively by the maximum observable level and by the fluorescence detector threshold at the beginning of the cascade development at the Pierre Auger Observatory.
In order to perform a composition study, we need a set of parameters sensitive to the primary mass: the discrimination among the different components is done using the well known fact that heavy primary induced showers develop faster in the atmosphere with respect to light induced ones (e.g. they reach the cascade maximum at smaller atmospheric depths), because of the higher nucleus-air cross section for showers of the same primary energy and zenith angles. We expected to extract a set of observables from the longitudinal curves, suitable for showing this behavior, hence introducing the following parameters:

- X_{max}, N_{max}: atmospheric depth of shower maximum and number of charged particles at shower maximum;

- $p10$, $p50$, $p90$: atmospheric depths at which the 10%, 50%, 90% of the whole integral profile are reached. These are sort of indicators about the "rise time "of the longitudinal profiles;

- $d10$, $d50$: derivative of the longitudinal profiles sampled at $X = p10$, $X = p50$. These represent observables correlated with the rapidity of the cascade development

towards its maximum;

- E, θ: primary energy and zenith angle (these are not directly correlated with the mass).

The numerical values of the first two parameters (X_{max} and N_{max}) were evaluated by fitting the simulated profiles $N_{ch}(X)$ for charged particles in the range 200-870 g/cm^2 with a standard 6-parameters Gaisser-Hillas function:

$$N_{ch}(X) = N_{max} \frac{X - X_0}{X_{max} - X_0}^{\frac{X_{max}-X_0}{a+bX+cX^2}} \exp\left(\frac{X_{max} - X}{a + bX + cX^2}\right) \tag{1}$$

The integral I of the whole profile in the above-mentioned range was evaluated by numerically integrating the profile curves, specified at a certain number of points (at least greater than 4), with a NAG routine, which evaluates the integral using a third-order finite-difference formula, according to a method due to Gill and Miller [14]. The integral between successive points is calculated by a four-points finite-difference formula centered on each interval, except in the case of the first and last intervals, where four-point forward and backward difference formulae respectively are employed. The values of the parameters $p10$, $p50$, $p90$ were then determined interpolating with a first-order polynomial in the interval, inside of which the required $10\%I$, $50\%I$, $90\%I$ integrals are reached. The choice of using such NAG routine is motivated by the fact that it does work with unequally-spaced points, as the points of the experimental profiles actually are.
The parameters $d10$ and $d50$ were determined taking the derivatives of the fit profile curves at $X = p10$ and $X = p50$.
The sensitivity to the primary mass for every chosen parameters was evaluated, using the simulated SetII, evaluating the following quantity at a given energy and zenith angle:

$$\eta = \frac{\overline{N}_{Fe} - \overline{N}_p}{\sqrt{[RMS_{Fe}]^2 + [RMS_p]^2}} \tag{2}$$

where $\overline{N}_{Fe}$ and $\overline{N}_p$ are the mean values of the distributions for every parameter, for fixed energies and zenith angles, for the proton and iron components, while RMS are the corresponding RMS values. η represents an estimate of the separability between the lightest and the heaviest mass components for different energies and zenith angles; larger η values correspond to a better discriminating power of the considered parameter. Analysis performed with SetII show a good discriminating power at all zenith angles, with smaller values of η at the increase of primary energy, for all the examined parameters.
The parameter space built in this way is therefore suitable for the neural network method application.

3 Neural network application to simulated data

This section presents the application of a neural network technique to the identification problem, describing the design of the network used, and the steps followed to perform the analysis.

3.1 The network design

A feed forward neural network (NN) is structured in parallel layers of neurons, connected to neurons in adjacent layers by weighted connections, indicating the strength of the neuron link. The input layer is connected to the input data vector and an indefinite number of hidden layers process the signal towards the output layer which returns the final response of the network to the presented input data. Figure 1 shows a typical architecture of a NN, deduced from our analysis, the design of which will be discussed.

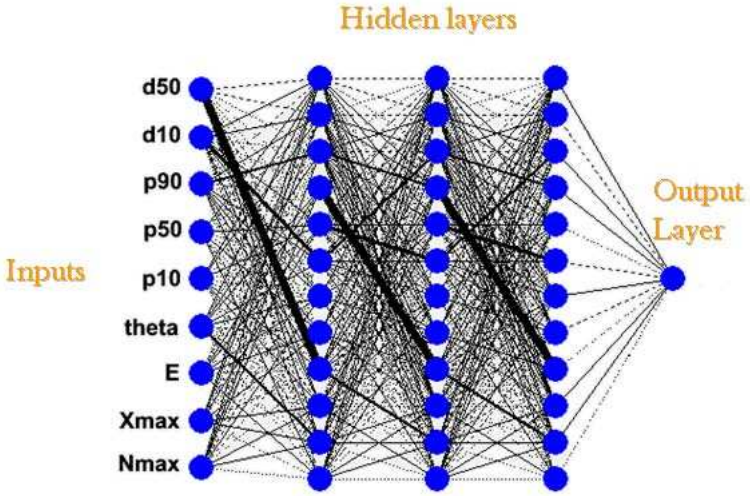

Figure 1: A typical neural network architecture, designed with the feature set discussed in section II as input parameter vector. The lines represent the neuron weights: larger line sizes indicate greater weight values.

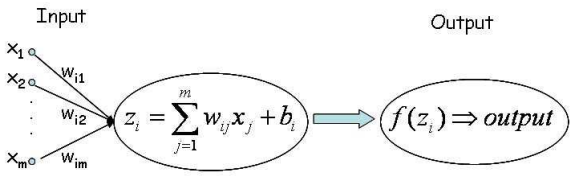

Figure 2: A single neuron from Figure 1 of index i: the input vector components x_i, the weights w_{ij} (j=1,...m), the biases b_i, the transfer function f and the output signal $f(z_i)$ are showed.

The basic processing unit in the network is the neuron: the input signals x_i (i=1,m) coming from each input channel are linearly transformed by applying a multiplicative weight w_{ij} and an additive bias b_i to form the net neuron input z_i:

$$z_i = \sum_{j=1}^{m} w_{ij} x_j + b_i \tag{3}$$

The neuron output is obtained by applying a transfer function $f(z_i)$ to the net input (see Figure 2). Common forms of such activation functions are the simple linear function $f(z_i) = \alpha z_i + \beta$, or the sigmoidal form functions, as well as the logistic function $f(z_i) = \frac{1}{1+\exp(-\alpha z_i)}$ and the hyperbolic tangent function $f(z_i) = \frac{\exp(z_i)-\exp(-z_i)}{\exp(z_i)+\exp(-z_i)}$.
After testing several network architectures, we obtained good results using a net with an input vector of dimension 9, 3 hidden layers, each one with 12 neurons, and an output layer with one neuron. The activation functions are logistic sigmoid in the hidden layers and linear in the output layer.
Next step is the choice of the training algorithm. The training data is a set of N events $(\mathbf{x}_i, y_i)$ $i = 1, ...N$, defined by the 9-dim input vector $\mathbf{x}_i \equiv$(X_{max}, N_{max}, E, θ, $p10$, $p50$, $p90$, $d10$, $d50$)$_i$ and by the desired output vector (the mass identity of the event) y_i. The supervised training algorithm minimizes the difference between the desired output y_i and the network computed output t_i, by adjusting iteratively the weights and biases of the net in order to minimize a given error function E. The error function used for the present analysis is the standard square error function:

$$E = \frac{1}{2} \sum_{i=1}^{N} [y_i(\mathbf{x}, \mathbf{w}) - t_i]^2 \tag{4}$$

Some backpropagation training algorithms have been tested (steepest descent, conjugated gradient and quasi-Newton algorithms). We achieved better identification performances with quasi-Newton methods, since other algorithms often return bad or local minima of the error function. We used a quasi-Newton algorithm with the Broyden-Fletcher-Goldfarb-Shanno (BFGS) error minimization formula [15].
Next subsection will describe the identification procedure we followed.

3.2 The identification method

The identification analysis proceeds as follows:

- *Pattern selection*: we divided SetI in two subsets in order to use them as network training set and testing data set. We stopped the network learning phase and evaluated the network performances using the latter set. To be more specific, a cross

validation set should be used to stop the training phase, and a further independent data set should be used to test the network efficiency. This will be done in a future analysis, as soon as a larger simulated data set will be available;

- *Feature pre-processing*: we normalized the features in the range [-1,1] to avoid large dynamics among the network inputs;

- *Training phase*: we trained the network to return a value of 0 or 1 in presence of a proton or iron event, respectively. The learning phase was stopped at a given epoch when the network began to show a clear overtraining behavior, corresponding to a loss of generality in the identification procedure, e.g. when the network error calculated over the test sample stopped to fall down and began to increase (see Figure 3).

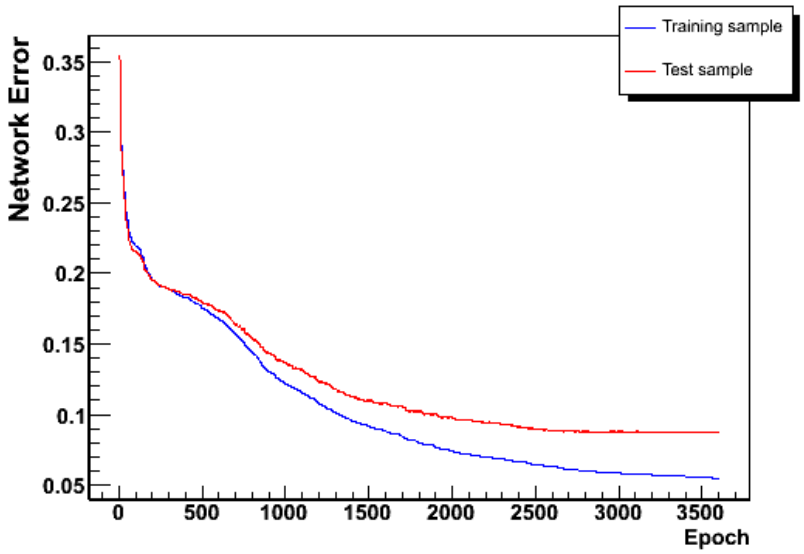

Figure 3: Example of network error trend as a function of the number of iterations performed during the learning phase, for the training sample (solid blue line) and for the test sample (solid red line).

- *Evaluation of the results*: we evaluated the performances of the method by means of the identification efficiency ε and purity P for a given mass class i of primaries :

$$\varepsilon^{(i)} = \frac{N^{(i)}_{right}}{N^{(i)}_{true}} \tag{5}$$

$$P^{(i)} = \frac{N^{(i)}_{right}}{N^{(i)}_{right} + \sum_{j \neq i} N^{(j)}_{wrong}} \tag{6}$$

where N_{true}, N_{right} and N_{wrong} represent the true number of events for the given mass class, the number of correctly identified events and the number of misclassified events. N_{right} was evaluated through a cut over the network output: events with an output smaller than 0.5 were recognized as protons, otherwise as iron nuclei.

4 Results

In this section we report the results of the classification analysis, in terms of identification efficiency and purity achieved.
Figure 4 shows the outputs computed by the net in presence of the training data set (on the left) and the test set (on the right). The blue histograms correspond to the true proton

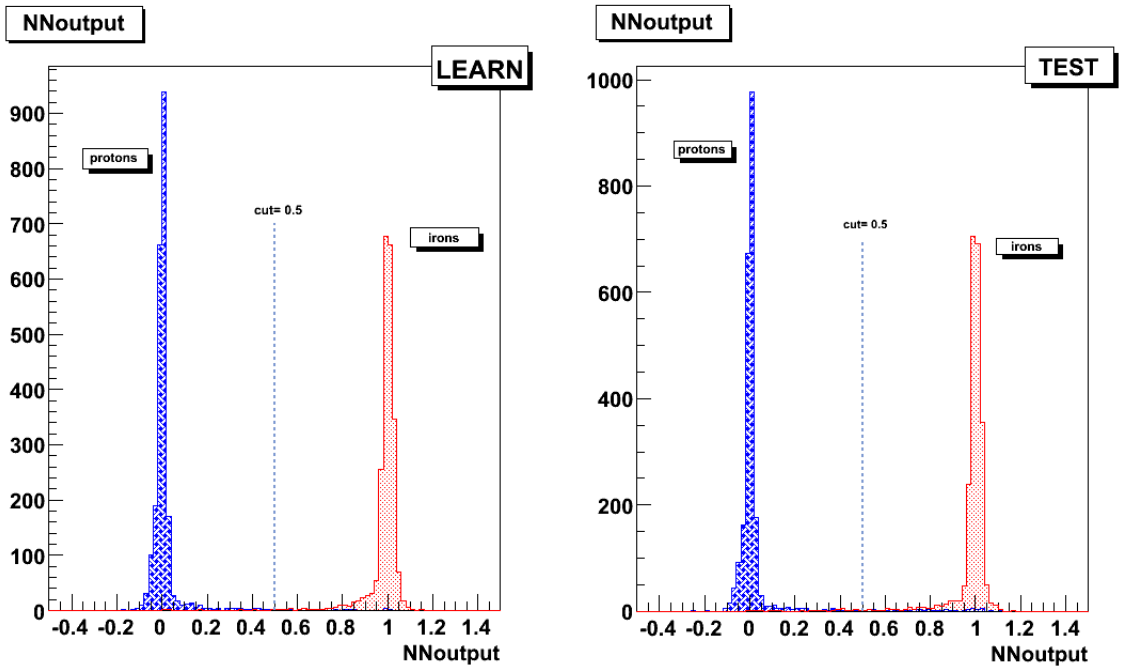

Figure 4: Output computed by the net in presence of the training data set (on the left) and the test set (on the right). The blue histograms correspond to the true proton events, while the red ones are the true iron events. The dashed line shows a cut at 0.5 in the net outputs to separate the two mass classes.

events, while the red ones are the true iron events. As we can clearly see, the net is able to associate the proton and iron events to the desired outputs with very little misclassifications. The identification efficiency and purity, relative to the chosen cut at 0.5, are shown in Table 1 for the proton and iron mass classes. We test the designed network also using SetII, which is a data set formed with fixed energy and zenith angles events with a better

Table 1: Identification efficiency, eq. (5), and purity, eq. (6), for the training and the test samples.

	LEARN		**TEST**	
	Efficiency	Purity	Efficiency	Purity
p	99.33%	99.69%	97.63%	98.56%
Fe	99.69%	99.34%	98.57%	97.64%

Table 2: Identification efficiency, eq. (5), and purity, eq. (6), for the training, TESTI and TESTII samples.

	LEARN		**TEST I**		**TEST II**	
	Efficiency	Purity	Efficiency	Purity	Efficiency	Purity
p	98.89%	99.46%	98.24%	98.58%	96.74%	98.20%
Fe	99.47%	98.90%	98.57%	98.23%	98.23%	96.79%

thinning level. In this case we noticed that the obtained performances degraded with respect to the ones showed in Table 1: the net seems to be unable to face off the intrinsic shower-to-shower fluctuations at a given primary energy and zenith angle.

The misclassifications are stronger at smaller zenith angles, because of a lack of statistics in the used training sample, due to the simulated zenith angle distribution of SetI. The net performances can be restored by including in the training sample a subset of SetII and re-executing the learning phase. The results, showed in Table 2, demonstrate that the efficiency and purity for the SetI and SetII events (denoted with TEST I and TESTII in Table 2) are basically the same.

We tested our method also in presence of a four components primary flux (proton, helium, oxygen, and iron) using the simulated sample SetII and assigning a desired net output of 0, 1, 2, 3, respectively, to the four classes. Results are showed in Figure 5. By cutting at 0.5, 1.5, 2.5 we separated the four classes, obtaining the classification matrix showed in Table 3.

The diagonal values are the identification efficiency of the four classes, while the non-diagonal elements give information about the misclassification of a class with respect to the others. Results show that the lightest and heavier components are better reconstructed, while a stronger contamination is found in the intermediate components.

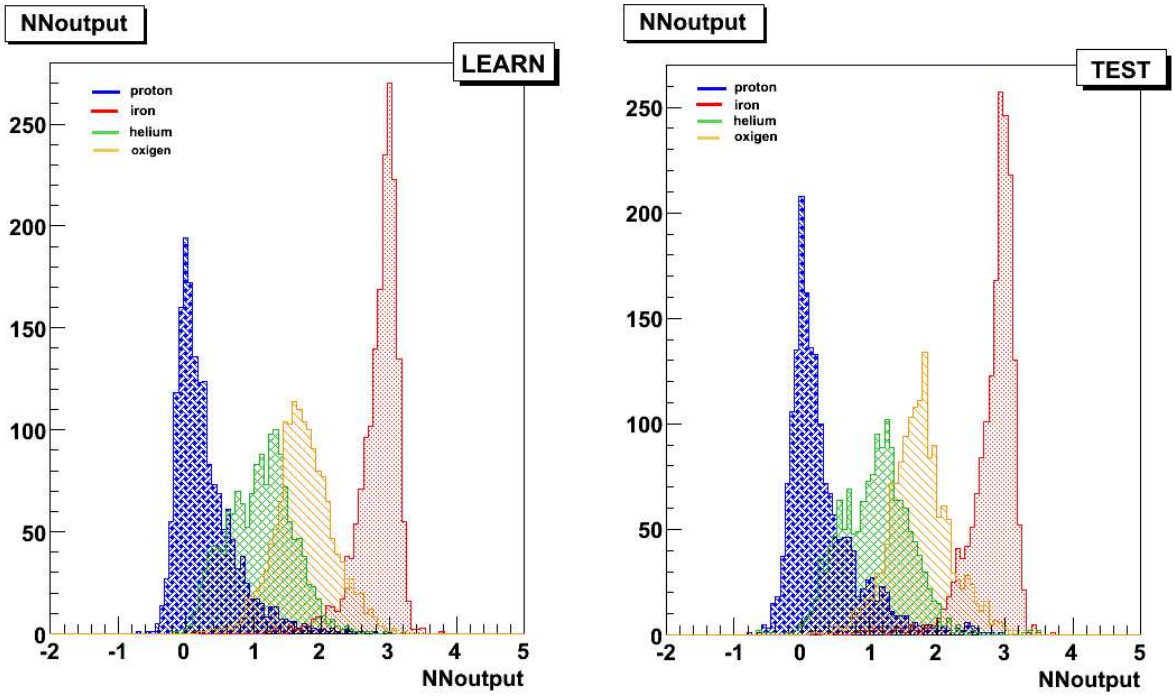

Figure 5: Output computed by the net in presence of the training data set (on the left) and the test set (on the right). The blue and red histograms correspond to the true proton and iron events, while the green and orange ones are the true helium and oxygen events.

5 Conclusion and future perspectives

We proposed and tested the neural network approach to the mass identification problem of high energy cosmic rays.

We studied mass discrimination in the case of CORSIKA simulated showers with a 2-components (proton and iron nuclei) and 4-components (proton and helium, oxygen, iron nuclei) mass flux, making use of parameters from the longitudinal profiles. In the first case we obtain excellent performances, with very small misidentification probabilities, of the order of 2%.

In the second case we found misclassification probabilities of 26%, 33%, 33% and 14% for the above mass classes, but these are obtained using fixed energy and zenith angle events.

The obtained results indicate that a better analysis should be performed using higher statistics and homogeneous data samples in the energy and zenith angles variables, since the network performances have been found to strongly depend from the used training data sets. We plan to perform a more accurate analysis, especially in the multi-component case, using CONEX [16][17] as shower simulation code and the latest version of QGSJET hadronic model, and to develop a general method able to determine also the mean com-

Table 3: Classification matrix and identification for the training and test samples.

	LEARN				
	Classification $P_{C_i \to C_j}$				**Purity**
	$C_j = p$	$C_j = He$	$C_j = O$	$C_j = Fe$	
$C_i = p$	76.29%	21.20%	2.23%	0.29%	87.25%
$C_i = He$	12.40%	68.47%	18.87%	0.27%	57.06%
$C_i = O$	0.53%	26.47%	67.93%	5.07%	66.91%
$C_i = Fe$	0.06%	0.29%	10.40%	89.26%	94.84%

	TEST				
	Classification $P_{C_i \to C_j}$				**Purity**
	$C_j = p$	$C_j = He$	$C_j = O$	$C_j = Fe$	
$C_i = p$	74.74%	21.60%	3.03%	0.63%	86.79%
$C_i = He$	12.53%	67.40%	19.53%	0.53%	55.86%
$C_i = O$	0.47%	26.87%	67.13%	5.53%	64.39%
$C_i = Fe$	0.23%	1.03%	12.06%	86.69%	93.70%

position or the energy spectra for the single mass components. CONEX uses a hybrid approach, based on the Montecarlo standard method and on the numerical integration of the cascade equations, making possible to produce simulated showers with smaller CPU times. A library of simulated showers for five mass components (proton and helium, oxygen, silicon, iron nuclei) is already available.
As future perspective, we plan to take into account the response of the FD detector at the Pierre Auger Observatory, evaluating the effects introduced by the detector over the used shower parameters, and to study the performances obtained with other hadronic interaction models, such as Sibyll or Nexus.

6 Acknowledgements

The authors thank Dr. M. Risse for the simulated cascades used in this work, and Prof. M. Russo for remarks and useful discussions.

7 References

References

[1] M.S. Longair, High Energy Astrophysics, Cambridge Univ. Press (1981).

[2] T. Antoni *et al*, Astropart. Phys. **24**, 1 (2005).

[3] G. Navarra, Nucl. Phys. B (Proc.Suppl) **136**, 265 (2004).

[4] G.Thomson, Nucl. Phys. B (Proc. Suppl) **136**, 28 (2004).

[5] S.P. Knurenko *et al*, Nucl. Phys. B (Proc. Suppl) **151**, 92 (2006).

[6] K. Shinozaki, Nucl. Phys. B (Proc. Suppl) **151**, 3 (2006).

[7] M.T. Dova *et al*, Astropart. Phys. **21**, 597 (2004).

[8] M. Ave *et al*, Astropart. Phys. **19**, 61 (2003).

[9] A.A. Watson, Nucl. Phys. B (Proc. Suppl) **151**, 83 (2006).

[10] B.R. Dawson et al., Astropart. Phys. **9**, 331 (1998).

[11] P. Mantsch for the Pierre Auger Collaboration at the 29th International Cosmic Ray Conference (ICRC 2005), Pune, India, 3-11 Aug 2005, astro-ph/0604114.

[12] D. Heck *et al*, FZKA Report Forschungszentrum Karlsruhe 6019 (1998).

[13] N.N. Kalmykov *et al*, Nucl. Phys. B (Proc. Suppl.) **52**, 17 (1997).

[14] P.E. Gill and G.F. Miller, Comput. J. **15**, 80 (1972).

[15] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press (1995).

[16] T. Pierog *et al*, Nucl. Phys. B (Proc. Suppl.) **151**, 159 (2006).

[17] N.N. Kalmykov *et al*, Astropart. Phys. **26**, 420 (2007).