PROCEEDINGS OF SCIENCE

Towards mass composition study with KASCADE using deep neural networks

M. Kuznetsov,[a] **N. Petrov,**[a,b,c] **I. Plokhikh**[a,b,d] **and V. Sotnikov**[a,e,*]

[a] *Institute for Nuclear Research of the Russian Academy of Sciences, 117312, Moscow, Russia*

[b] *Novosibirsk State University, 630090, Novosibirsk, Russia*

[c] *Budker Institute of Nuclear Physics, 630090, Novosibirsk, Russia*

[d] *Institute of Thermophysics SB RAS, 630090, Novosibirsk, Russia*

[e] *JetBrains Limited, 2409 Nicosia, Cyprus*

E-mail: vladimir.sotnikov@jetbrains.com

We present new insight into the ongoing machine learning analysis of KASCADE experiment archival data, that contain air shower events with $\sim 1 - 100$ PeV primary energy. The aim of the study is to improve the accuracy of high-energy cosmic rays mass composition reconstruction with respect to the standard KASCADE technique. We introduce five mass groups: protons, helium, carbon, silicon and iron nuclei and interpret the reconstruction process as a classification task. We employ a random forest technique as well as two promising neural network architectures — a self-attention perceptron and a convolutional neural network. These models are being trained with KASCADE CORSIKA simulations. We examine the behavior of the mass composition reconstruction for several hadronic interaction models and additionally check the credibility of our methods with a small "unblinded" part of the real KASCADE data.

**** 27th European Cosmic Ray Symposium - ECRS ****
**** 25-29 July 2022 ****
**** Nijmegen, the Netherlands ****

*Speaker

PoS(ECRS)092

1. Introduction

High energy cosmic rays (CRs) — charged particles that are reaching the Earth from space were studied extensively during last several decades. Despite the significant improvement in both experiment and simulation techniques, important questions of their origin and behaviour remain unanswered (see e.g. [1, 2] for a review). In particular, the transition between CRs of galactic and extragalactic origin is expected to occur somewhere in a three-decade wide interval between the "knee" and the "ankle" of their spectrum. But the precise position of the transition region is unclear as well as the CR mass composition around this region. The knowledge of the latter would shed light on the nature of CR sources and the mechanism of CR production. The absence of the answers to these questions for such a long time is due to the several interrelated problems: the reconstruction of mass and charge of CRs is complicated because of their indirect detection (via air-showers), the identification of the sources of CRs is complicated as well because of a poor knowledge of their charge and the value of galactic magnetic field. Moreover, different models of hadronic interactions give significantly different predictions for the same particles.

In these proceedings we present the results of a new analysis of KASCADE experiment archival data [3], that are provided by the KCDC service [4]. The analysis aims at a better separation of individual primary mass groups with a help of machine-learning techniques. Comparing to our previous study presented in Ref. [5], where only the random forest method was used, in this study we also employ convolutional neural net and self-attention perceptron. We estimate the accuracy of the methods and its dependence on the hadronic interaction models using the KASCADE Monte-Carlo. We also test the viability of the method with a small "unblinded" part of the real data. The neural nets show somewhat better mass-groups separation power than the random forest.

The outline of the text is the following: we briefly describe the MC and data we use in section 2 and the machine learning methods we employ — in section 3. In section 4 we estimate the performance of our methods with Monte-Carlo and unblinded data and conclude in section 5.

2. Data

We use KASCADE pre-selection data sets. The full archive consists of ~ 300 M air shower events in energy range $\sim 1 - 100$ PeV, detected by a 16x16 array of scintillating detectors during experiment operation from 1996 to 2013. The events contain the following reconstructed air shower features: energy E, shower core coordinates (x, y), arrival direction (zenith angle θ, azimuthal angle ϕ), muon and electron numbers ($\log_{10} N_e$, $\log_{10} N_\mu$) and shower age (s); raw readings from e/γ and μ detector stations and arrival times that represent a time stamp of first particle hitting the detector for a given event.

The KCDC [4] service provides CORSIKA [6] simulations of air-shower events generated for five CR primary mass groups: *H*, *He*, *C*, *Si* and *Fe*. These simulations provide the same properties as in the real data, reconstructed using the actual detector response. An example of the KASCADE experiment event is shown in Fig. 1. The detectors energy deposits in data sets are presented as 16x16 arrays. In this study we use only an integral signal of each detector for a given event but do not use a timing of this signal.

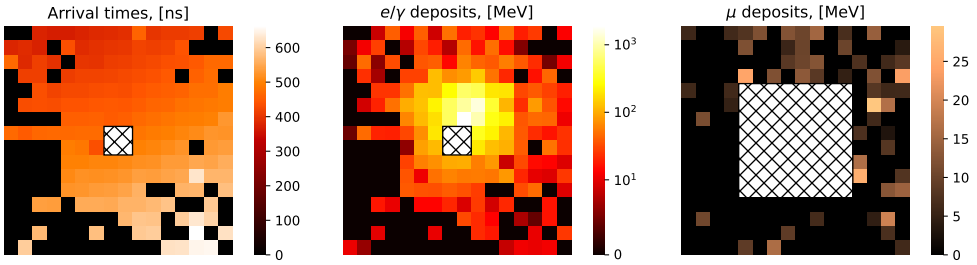


Figure 1: Example of the experimental event in the dataset. The matrices of arrival times, e/γ and μ deposits are shown. Reconstructed features of the event: $\log_{10}(E/\text{eV}) = 15.45$, $\theta = 19.37°$. Note, KASCADE does not have detector stations in the central 2x2 part for arrival times and e/γ deposits and in central 8x8 part for μ deposits. These areas are represented in the figure as unresponsive.

In this research we use the quality cuts recommended by KASCADE collaboration [7]: $\theta < 18°$, $x^2 + y^2 < 91$ m, $\log_{10} N_e > 4.8$, $\log_{10} N_\mu > 3.6$. We use the cut on the shower age set by KCDC: $0.2 < s < 1.48$, it is somewhat tighter than the original KASCADE cut ($0.2 < s < 2.1$).

3. Machine learning methods

We use several different machine learning (ML) methods for event-by-event mass group classification. We are starting with a Random Forest (RF). This is the classical ML approach, which is our baseline. In our implementation it takes as input only reconstructed features of the event (energy, zenith angle, etc.) provided by KCDC. To incorporate more data to the analysis we use neural networks. These models accept energy deposits from e/γ and μ detector stations. In particular, we design a convolutional neural network (CNN) and self-attention multi-layer perceptron (MLP).

All the models are trained with three data sets, corresponding to three modern hadronic interaction models: QGSJet-II.04 [8], EPOS-LHC [9] and Sybill 2.3c [10].

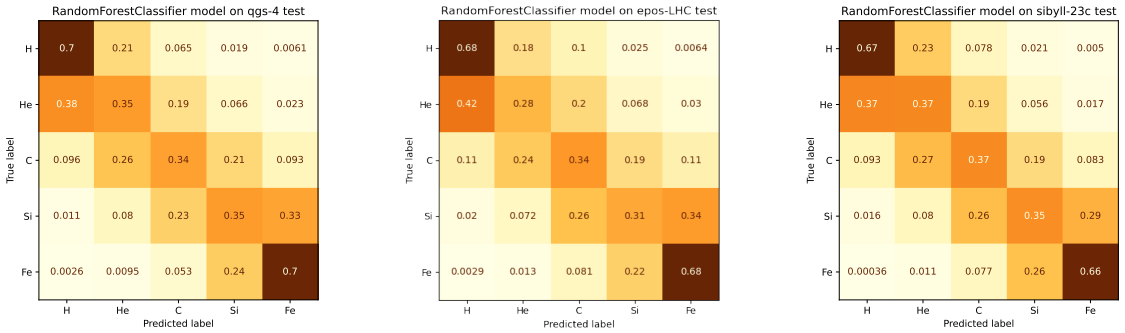
The data sets are divided into train, validation and test ones. Throughout the paper, we use train and validation sets for model's training (validation set is used for early stopping to avoid overfitting and hyperparameter tuning of the models). All the obtained models are evaluated on test sets of the corresponding data sets.

3.1 Random Forest

At first, we train the Random Forest classifier, it predicts the particle type directly for individual events for five mass groups. The following reconstructed features of the air shower were used as input parameters: E, x, y, θ, ϕ, $\log_{10} N_e$, $\log_{10} N_\mu$, s. The model was implemented and trained in scikit-learn package [11]. The hyper-parameters of the model were optimized using Grid Search algorithm.

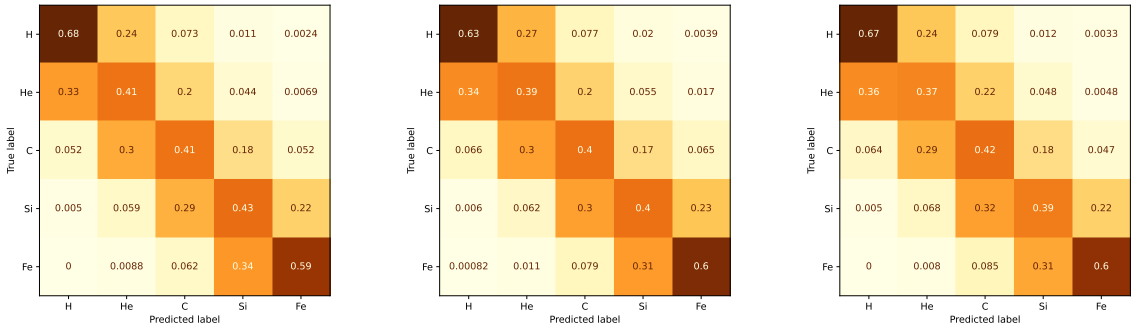
3.2 Convolutional neural network

Next, we developed a convolutional neural network for event-by-event classification of mass groups. The architecture of the presented model is similar to LeNet-5 [12], with integral signals from e/γ and μ detector stations for each event as input. Also, we append to the first dense layer



(a) **(b)** **(c)**

Figure 2: Confusion matrices for Random Forest Classifier models trained using three different hadronic interaction models (a) QGSJet-II.04, (b) EPOS-LHC and (c) Sybill 2.3c. True mass groups are normalized per unit.



(a) **(b)** **(c)**

Figure 3: Confusion matrices for CNN models trained using three different hadronic interaction models (a) QGSJet-II.04, (b) EPOS-LHC and (c) Sybill 2.3c. True mass groups are normalized per unit.

of the model the following reconstructed features: $\log_{10} N_e$, $\log_{10} N_\mu$, s. It helps to speed up the convergence of the model. This classifier is implemented in PyTorch [13]. The model has $\sim 30\,000$ trainable parameters.

3.3 Self-attention multi-layer perceptron

The last model is MLP. The key feature of the model is the self-attention layers [14]. This model is also trained for event-by-event classification of mass groups. The inputs of the MLP are integral signals from e/γ and μ detector stations only. This classifier is implemented in TensorFlow [15]. It has $\sim 36\,000$ trainable parameters.

4. Results

The confusion matrices of the Random Forest classifier, CNN and Self-attention perceptron are shown in Fig. 2, 3, 4 respectively. For all models, we see a diagonal structure of the confusion

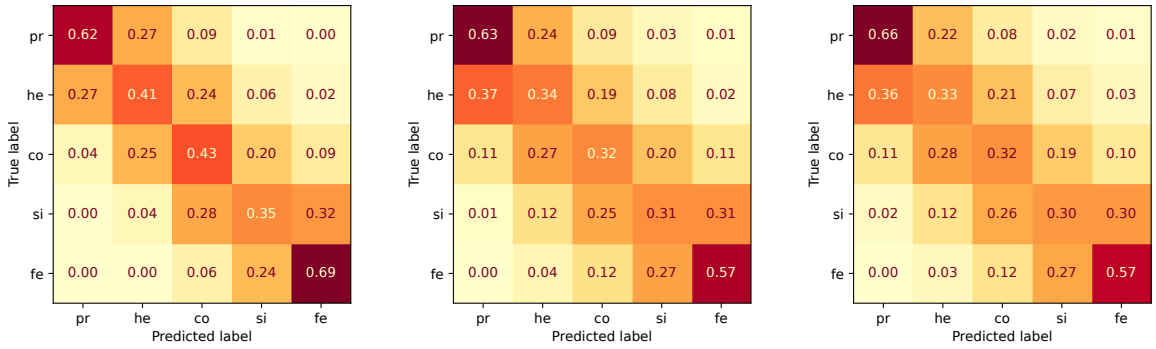


Figure 4: Confusion matrices for MLP models trained using three different hadronic interaction models (a) QGSJet-II.04, (b) EPOS-LHC and (c) Sybill 2.3c. True mass groups are normalized per unit.

Mass group	RF Classifier	CNN	MLP
H	3.67%	1.99%	2.49%
He	2.96%	2.52%	2.87%
C	3.11%	2.58%	6.05%
Si	3.25%	3.52%	3.14%
Fe	3.51%	3.84%	3.65%

Table 1: MAE between true and predicted fractions of mass groups for different models. All the results are performed for QGSJet-II.04 hadronic interaction model.

matrices. The best discrimination is observed for the lightest and the heaviest mass groups of particles (*H*, *Fe*).

To estimate the uncertainty of the method for each mass component we computed the mean absolute error (MAE) between its true fraction in ensembles and the predicted one. For this purpose we create 2000 random ensembles of 5000 events each using the test set.

The obtained results are shown in Table 1. For the sake of compactness we show only the results for QGSJet-II.04 hadronic interaction model. As can be seen from the table, most MAEs are in the range of about 2 – 4%. Only the MAE of the *C* component for MLP model is at 6% level.

In Fig. 5 we show the true spectra (they set to $\sim E^{-2.7}$ in KCDC Monte-Carlo) and the reconstructed spectra of separate mass components for QGSJet-II.04 model Monte-Carlo and CNN reconstruction. One can see, that the NN reconstruction does not lead to significant distortions of the spectra and satisfactory predicts the mass composition. We did not perform the unfolding corrections for the reconstruction of these spectra and leave this for further studies.

To test our methods for possible hidden systematics we use a semi-blind data analysis in this study. Namely, we divided KASCADE experimental data into blind and unblind parts at a ratio of 80%:20%. The blind data part is not used here and left for full-scale composition study. While the unblind part is used for the test of the composition reconstruction method. In Fig. 6 we shown the spectra for the separate mass components of the unblind part of the KASCADE data according to the reconstruction of CNN model trained with QGSJet-II.04 Monte-Carlo. One can see that the

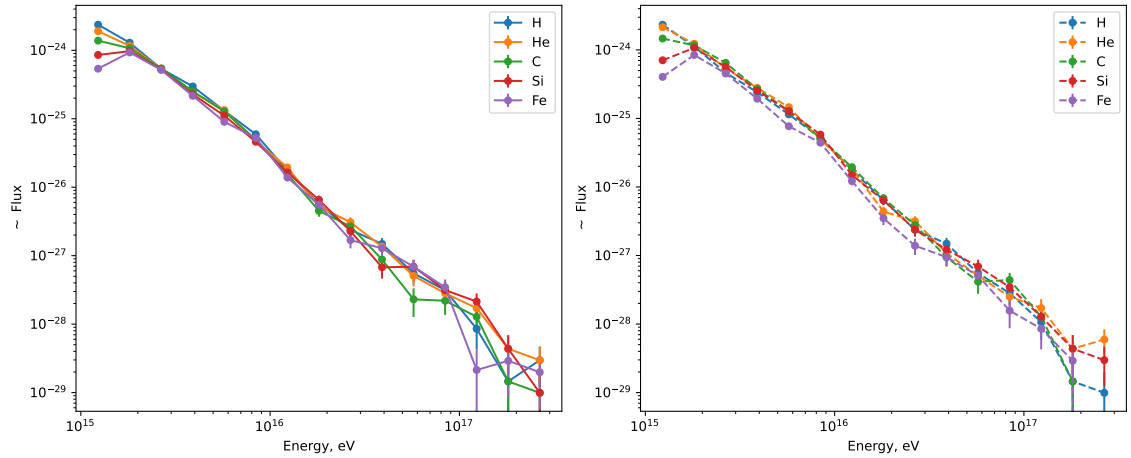


Figure 5: Spectra for the CNN model trained using QGSJet-II.04 hadronic interaction model. The figures represent true (left) and reconstructed (right) spectra by the model from the test part of QGSJet-II.04 simulations dataset.

spectra of separate mass groups are smooth at lower energies where the statistical errors are small enough. This behaviour of the spectra increases the credibility of our composition analysis method. At the same time the relation of different mass groups spectra is different than that reconstructed with the standard KASCADE methods [16], although that reconstruction was based on the older hadronic interaction model QGSJet-II.02. We also need to note that the usage of the full data set and the unfolding methods would change this picture. Therefore, we left this issue for a future analysis.

5. Conclusion

We have presented the updated results of a mass composition analysis of the KASCADE air shower experiment archival data provided by the KASCADE Cosmic ray Data Center. We have trained and implemented the modern machine learning models to the KCDC data and Monte-Carlo and built the new event-by-event mass groups classifiers. It was shown that the accuracy of the neural nets classifiers are in general higher than that of the random forest method. The accuracy of the estimation of the separate mass group prediction within a mixed composition data set was performed and was found reasonable, of order few percent. The application of the method to the unblind part of the experimental data reveals no unexpected artifacts, therefore providing an additional justification of our reconstruction.

Acknowledgments

The work is supported by the Russian Science Foundation grant 22-22-00883. We appreciate the contribution on the initial stages of this project made by our colleagues: Dmitry Kostunin, Vladimir Lenok and Victoria Tokareva.

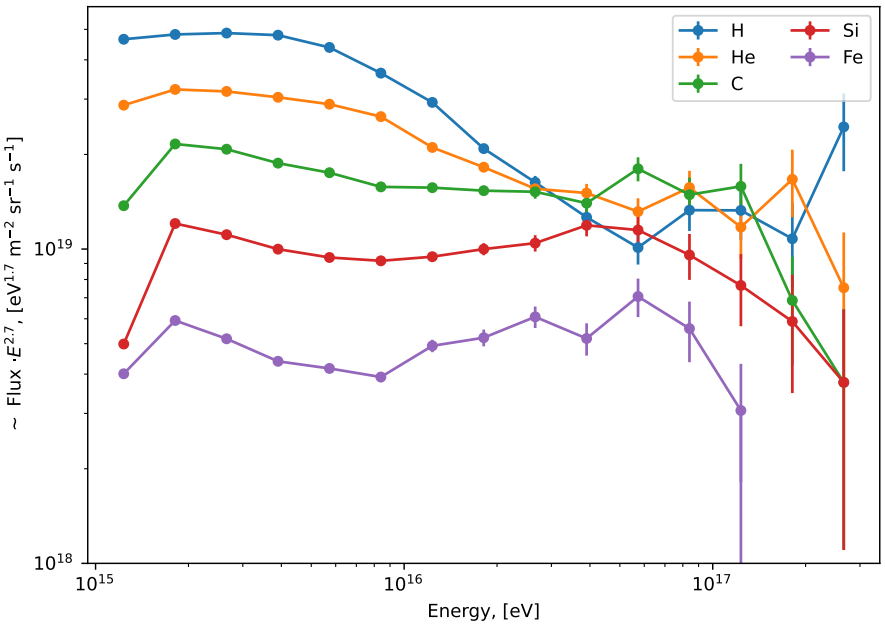


Figure 6: Mass composition spectra on the unblind experimental data (20% of the whole dataset) for the CNN model trained using QGSJet-II.04 hadronic interaction model.

References

[1] S. Gabici, C. Evoli, D. Gaggero, P. Lipari, P. Mertsch, E. Orlando et al., *The origin of Galactic cosmic rays: challenges to the standard paradigm*, *Int. J. Mod. Phys. D* **28** (2019) 1930022 [1903.11584].

[2] R. Alves Batista et al., *Open Questions in Cosmic-Ray Research at Ultrahigh Energies*, *Front. Astron. Space Sci.* **6** (2019) 23 [1903.06714].

[3] KASCADE collaboration, *The Cosmic ray experiment KASCADE*, *Nucl. Instrum. Meth. A* **513** (2003) 490.

[4] A. Haungs et al., *The KASCADE Cosmic-ray Data Centre KCDC: Granting Open Access to Astroparticle Physics Research Data*, *Eur. Phys. J. C* **78** (2018) 741 [1806.05493].

[5] D. Kostunin, I. Plokhikh, M. Ahlers, V. Tokareva, V. Lenok, P.A. Bezyazeekov et al., *New insights from old cosmic rays: A novel analysis of archival KASCADE data*, *PoS* **ICRC2021** (2021) 319 [2108.03407].

[6] D. Heck, J. Knapp, J. Capdevielle, G. Schatz, T. Thouw et al., *Corsika: A monte carlo code to simulate extensive air showers*, *Report fzka* **6019** (1998) .

[7] KASCADE collaboration, *KASCADE measurements of energy spectra for elemental groups of cosmic rays: Results and open problems*, *Astropart. Phys.* **24** (2005) 1 [astro-ph/0505413].

[8] S. Ostapchenko, *QGSJET-II: towards reliable description of very high energy hadronic interactions*, *Nuclear Physics B-Proceedings Supplements* **151** (2006) 143.

[9] T. Pierog, I. Karpenko, J.M. Katzy, E. Yatsenko and K. Werner, *EPOS LHC: Test of collective hadronization with data measured at the CERN Large Hadron Collider*, *Phys. Rev. C* **92** (2015) 034906 [1306.0121].

[10] F. Riehn, R. Engel, A. Fedynitch, T.K. Gaisser and T. Stanev, *Charm production in SIBYLL*, *EPJ Web Conf.* **99** (2015) 12001 [1502.06353].

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel et al., *Scikit-learn: Machine learning in Python*, *Journal of Machine Learning Research* **12** (2011) 2825.

[12] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, *Gradient-based learning applied to document recognition*, *Proceedings of the IEEE* **86** (1998) 2278.

[13] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan et al., *Pytorch: An imperative style, high-performance deep learning library*, in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox and R. Garnett, eds., pp. 8024–8035, Curran Associates, Inc. (2019).

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez et al., *Attention is all you need*, 2017, https://arxiv.org/pdf/1706.03762.pdf.

[15] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro et al., *TensorFlow: Large-scale machine learning on heterogeneous systems*, 2015.

[16] KASCADE collaboration, *Energy Spectra of Elemental Groups of Cosmic Rays: Update on the KASCADE Unfolding Analysis*, *Astropart. Phys.* **31** (2009) 86 [0812.0322].