Techniques for data analysis and primary mass reconstruction in the ENDA experiment

Kyrinov K. O., Kuleshov D. A., Stenkin Yu. V., Shchegolev O. B.

Institute for Nuclear Research, Russian Academy of Sciences

ISCRA-2023
June 27-29, 2023

Outline

ENDA

The Electron–Neutron Detector Array (ENDA) is being created in China within the large high-altitude air shower observatory (LHAASO) project.

Some ENDA-INR parameters: Digitization step - 32 nsec, detector trigger threshold - 3 mV, distance between detectors - 5 m.

Figure 1: Installation ENDA-LHAASO

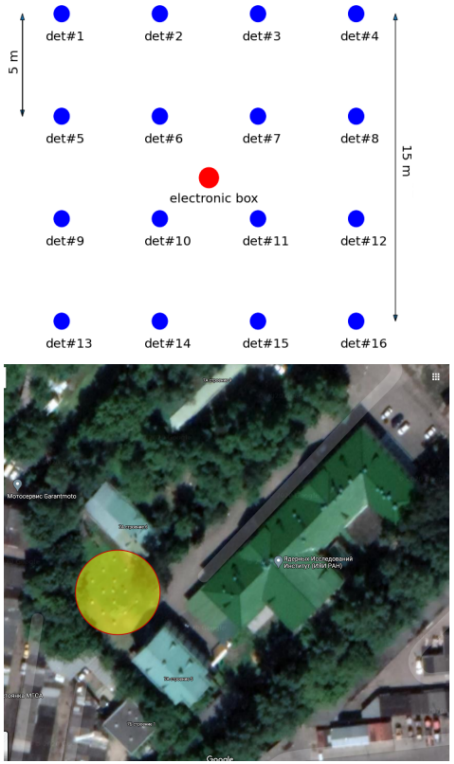


Figure 2: ENDA-INR configuration and installation location

Direction and core reconstruction

Standard approach: Maximum Likelihood Estimation with NKG parametrization for LDF (see appendix) and approximation of flat shower front for direction reconstruction.
ML: Convolutional Neural Nets or Artificial Neural Nets with different target variables parametrization. Features - energy deposit, response times, number of neutrons in each detector.

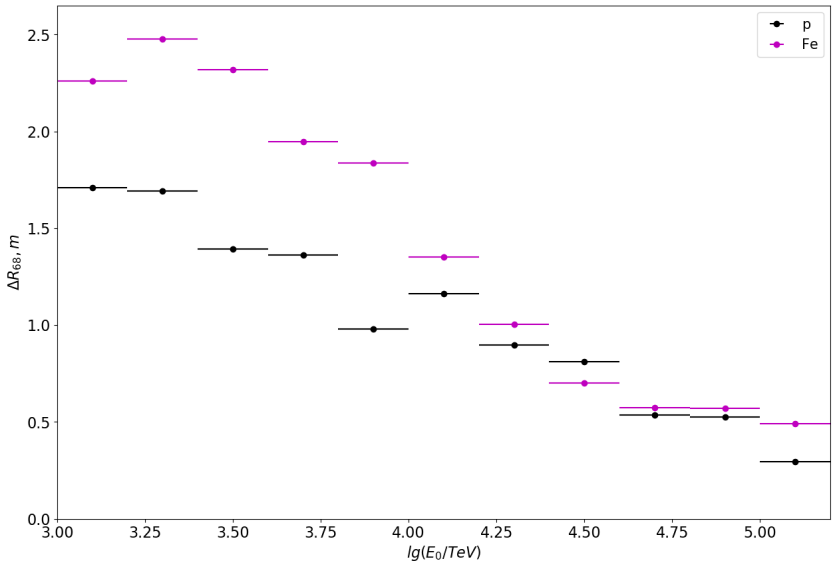


Figure 3: Core resolution for different primaries

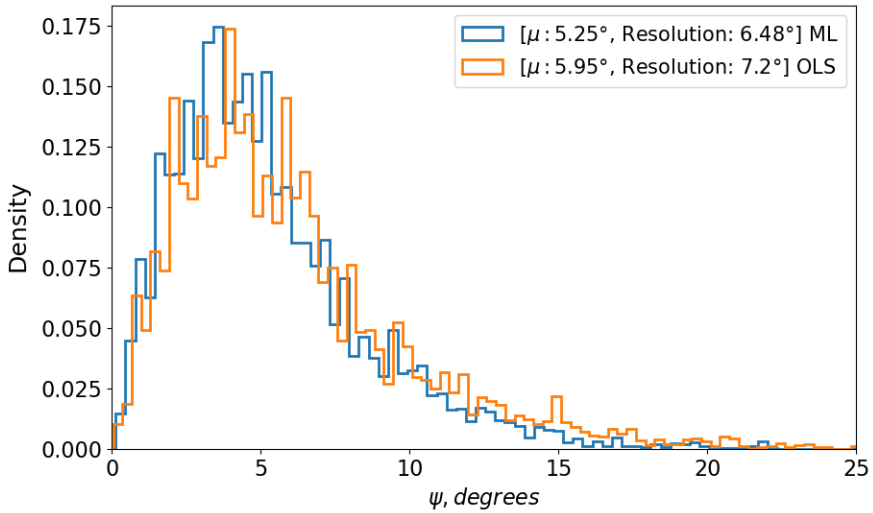


Figure 4: Distribution by the angle of deviation of the guiding vector of the arrival of the EAS

Standard approach:

$lg(E_0^{Rec.}) = A \times lg(N_e) + C$ **or** $lg(E_0^{Rec.}) = B \times lg(N_n) + D$ **or** Linear regression with polynomial features.

ML approach:

Algorithm: GBDT regression with XGBoost[a] and ANN.
Features: $[\theta, lgN_e, N_n, s,$ number of triggered detectors, $Q_{max}, R_{from\ center}]$ **or** $[\theta, lg\Sigma\rho, N_n, Q_{max}, R_{from\ center},$ number of triggered detectors$]$.

MC Simulation: balanced dataset with proton and iron primaries.
$E_0 \in [1 \div 300]$ PeV with differential slope -2.7.
Selection Cuts: $\theta < 30^\circ$, $lgN_e > 5$, number of triggered detectors > 7

Train set: Valid set: Test set = 3:1:1

[a] Chen T., Guestrin C. Xgboost: A scalable tree boosting system //Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.

Energy reconstruction

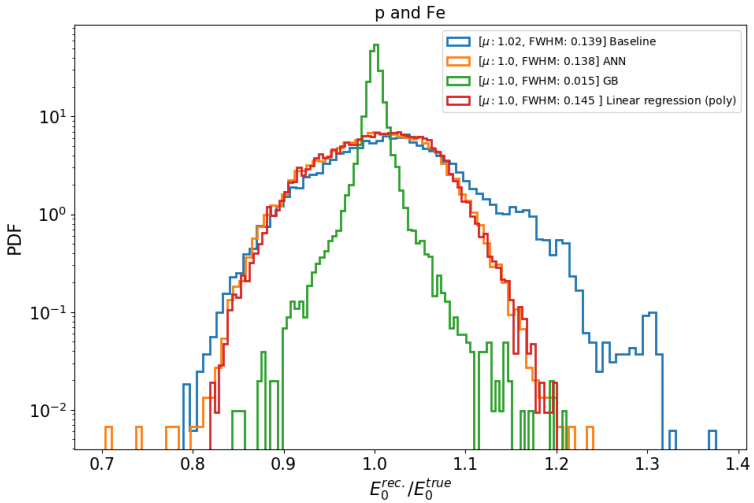


Figure 5: Distribution of $E_0^{rec.}/E_0^{true}$

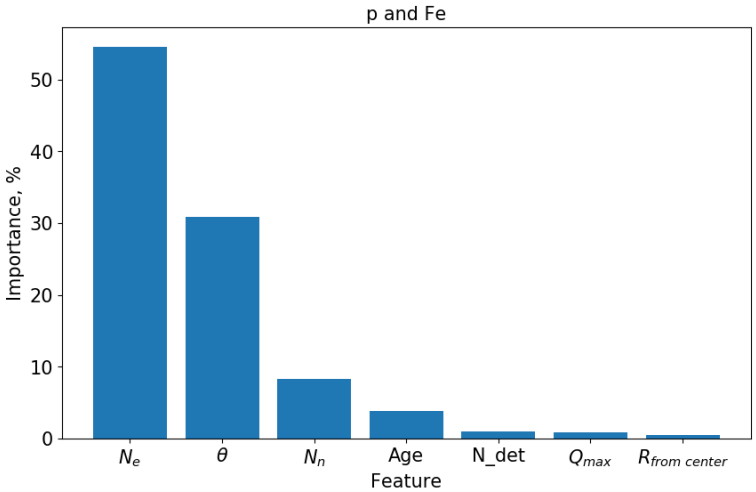


Figure 7: Feature importance

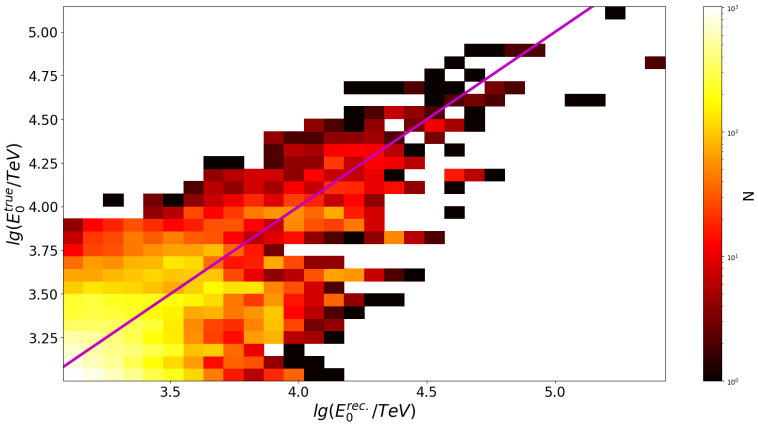


Figure 6: $E_0^{rec.}$ *vs* E_0^{true} for standard approach

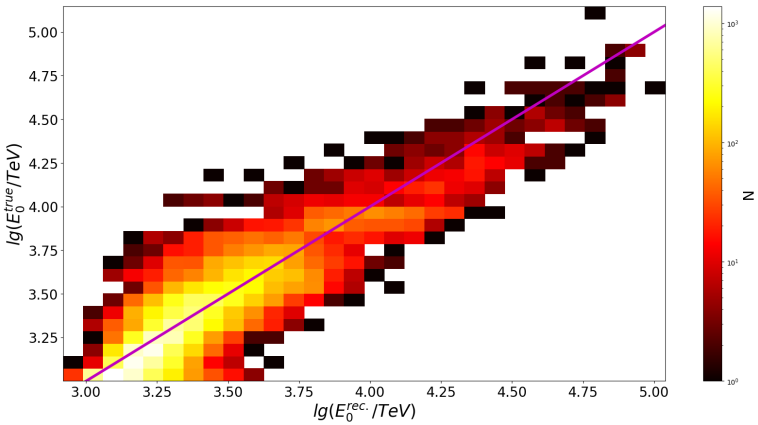


Figure 8: $E_0^{rec.}$ *vs* E_0^{true} for GBDT

Standard approach: Events with large fluctuations in the development of a shower can lead to the inapplicability of the selected a priori distribution in the likelihood function.

ML approach: Neural network is a "black box". In classification task softmax probability doesn't estimate uncertainty[b].

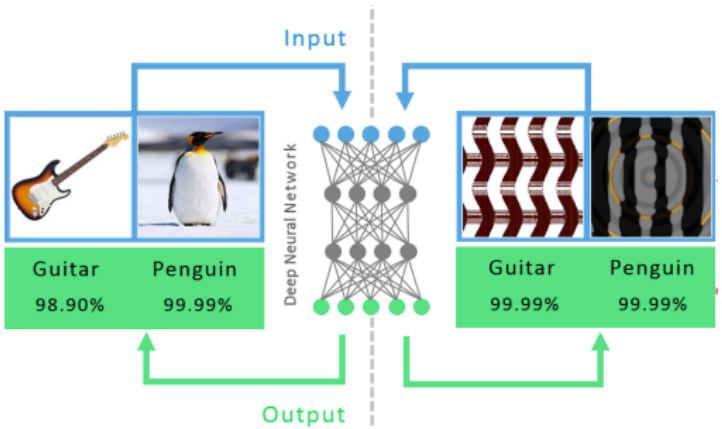
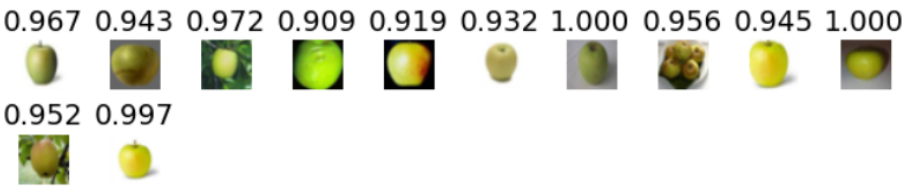


Figure 9: Deep neural network determined unrecognizable images as familiar objects with high confidence



CIFAR-100's *apple* misclassified as CIFAR-10's *frog* class with $p > 0.9$.

[b] Nguyen A., et al. "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

Two types of uncertainty

It is possible to divide the types of uncertainty into two classes. First called *data* uncertainty, which arises due to inherent class overlap or noise in the data. And uncertainty due to the model's inherent lack of knowledge about inputs from regions either far from the training data or sparsely covered by it, called *knowledge* uncertainty[c,d].
To estimate the knowledge uncertainty, it is necessary to use Bayesian methods and calculate the a posteriori distribution of the reconstructed parameters.

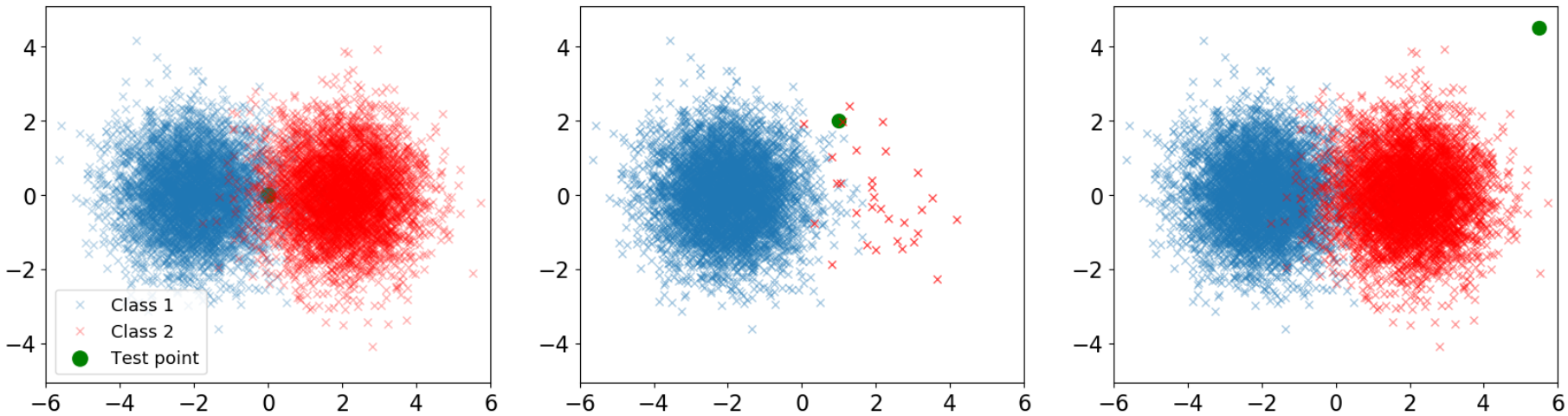


Figure 10: Data uncertainty (left), knowledge uncertainty (center and right)

[c] Yarin Gal, Uncertainty in Deep Learning, Ph.D. thesis, University of Cambridge, 2016.

[d] Andrey Malinin, Uncertainty Estimation in Deep Learning with application to Spoken Language Assessment, Ph.D. thesis, University of Cambridge, 2019.

Uncertainty estimation (Standard approach)

Rao-Kramer inequality

$$D(\hat{\theta}) \geq |j(\theta)|^{-1},$$

where $j(\theta) = -\mathbb{E}(\frac{\partial^2 ln\mathcal{L}}{\partial\theta_i \partial\theta_j}) - \textit{Fisher's information}, \mathcal{L} - \textit{Likelihood function}.$

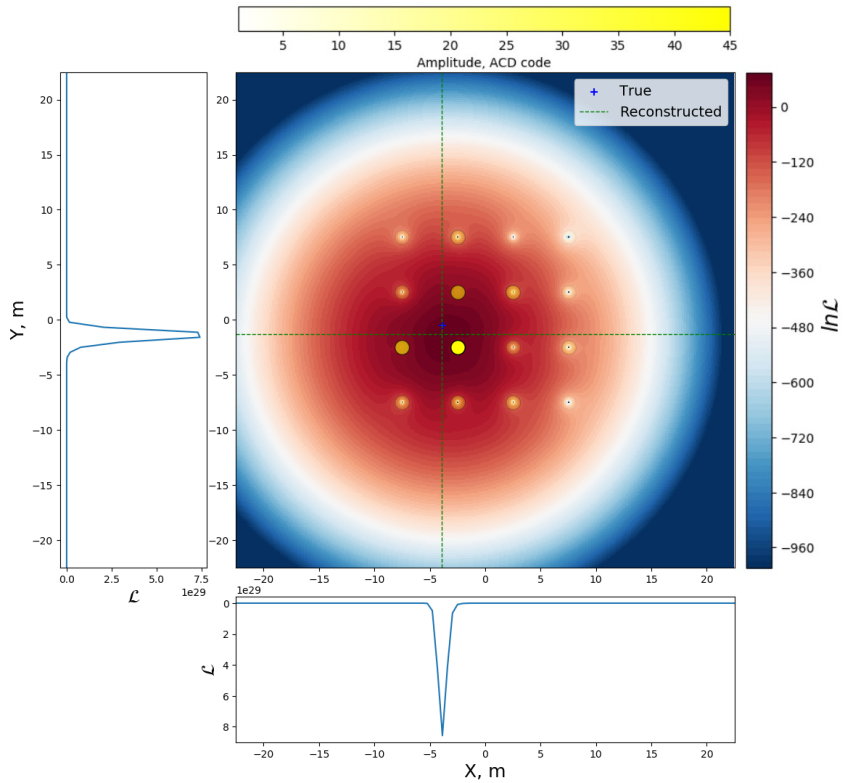


Figure 11: Event with a small ΔR error

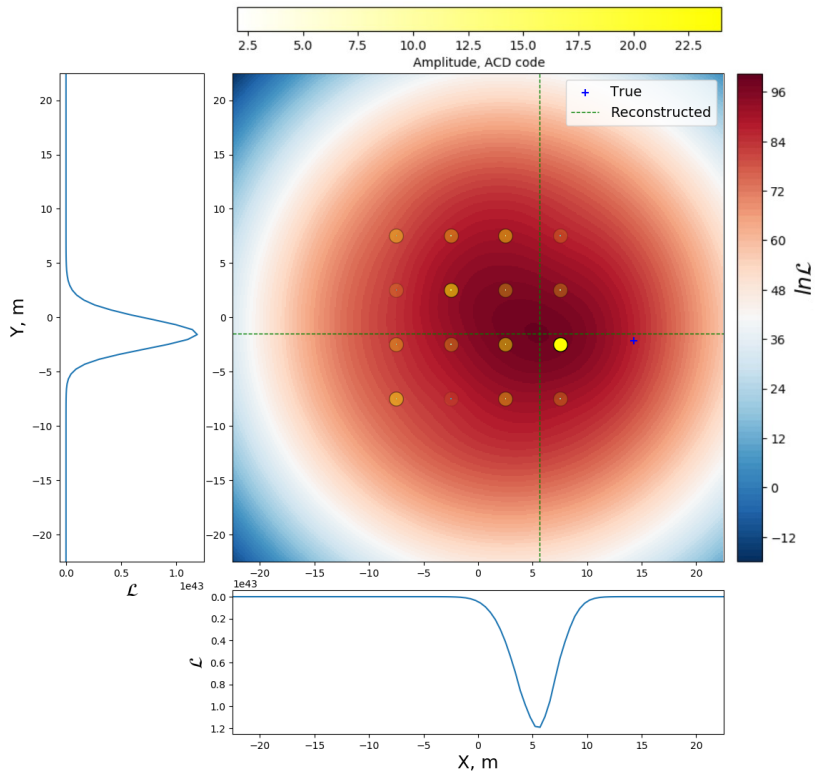


Figure 12: Event with a big ΔR error

For regression model use Gaussian Negative Log Likelihood (GNLL) loss and predict mean and variance of a normal distribution over the target variable y for a given feature input.

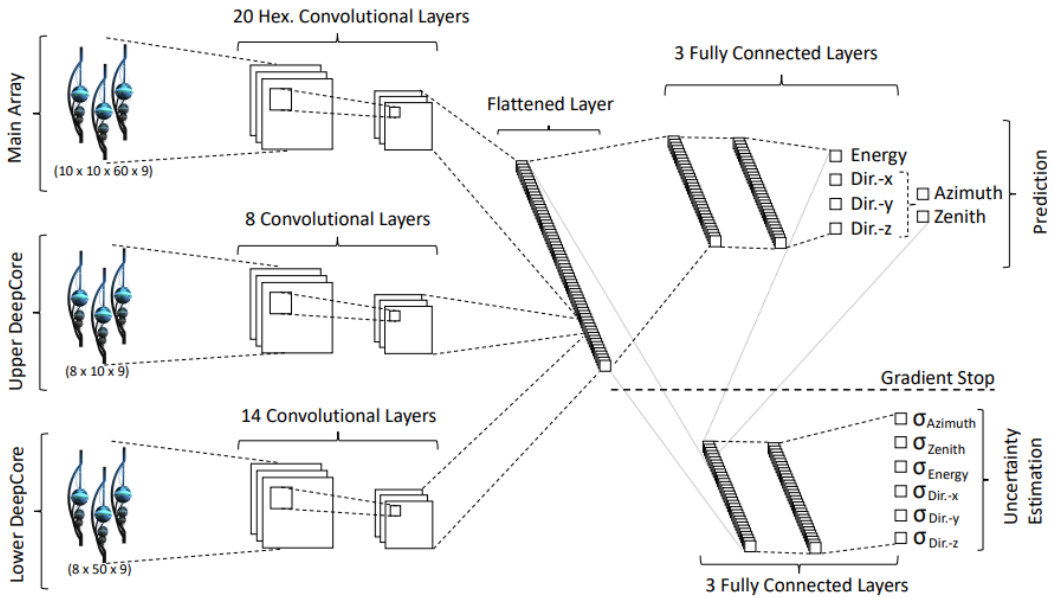


Figure 13: A neural network architecture for IceCube[a].

GNLL

$$\mathcal{L} = ln(\sqrt{2\pi\sigma^2}) + \frac{(y-\mu)^2}{2\sigma^2},$$

where y - target variable,
μ, σ^2 – *model prediction*.

Disadvantage: Capture only *data uncertainty* and doesn't capture *knowledge uncertainty*.

[a] Abbasi, R., et al. (2021). A convolutional neural network based cascade reconstruction for the IceCube Neutrino Observatory.

The approach is based on creating an independent ensemble of models, where each model predicts its own mean and variance. Uncertainty in predictions caused by uncertainty of knowledge and expressed as the level of dispersion between models in the ensemble.

$$p(\theta, \mathcal{D}) \rightarrow \begin{matrix} \text{model}^{(1)} \rightarrow \left(\bar{\theta}^{(1)}, \sigma^2_{\theta^{(1)}}\right) \\ \vdots \\ \text{model}^{(2)} \rightarrow \left(\bar{\theta}^{(2)}, \sigma^2_{\theta^{(2)}}\right) \\ \vdots \\ \text{model}^{(N)} \rightarrow \left(\bar{\theta}^{(N)}, \sigma^2_{\theta^{(N)}}\right) \end{matrix}$$

$$output \rightarrow \mathbb{E}\left[\bar{\theta}^{(i)}\right]$$

$$Data\ uncertainty \rightarrow \mathbb{E}\left[\sigma^2_{\theta^{(i)}}\right]$$

$$Knowledge\ uncertainty \rightarrow D\left[\bar{\theta}^{(i)}\right]$$

$$\mathcal{D} = \{x_i, y_i\}_{i=1}^{N}$$

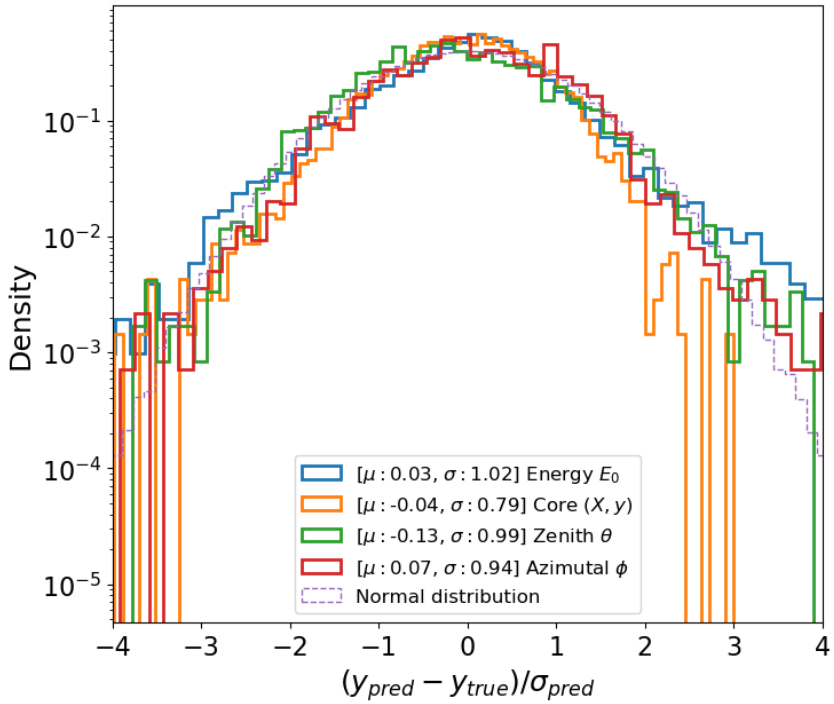


Figure 14: Data uncertainty estimation for reconstructed parameters in compare with $\mathcal{N}(0, 1)$

Features for primary particle identification

Task: Separation of proton showers from all others (He, N, Fe). Classes are balanced.

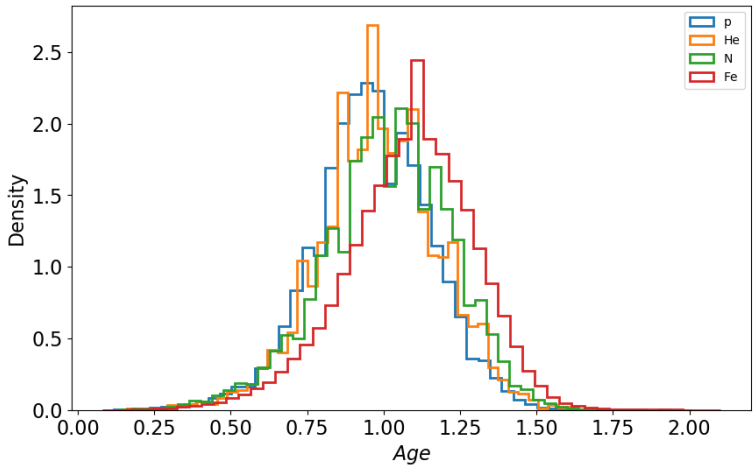


Figure 15: Distribution of Age

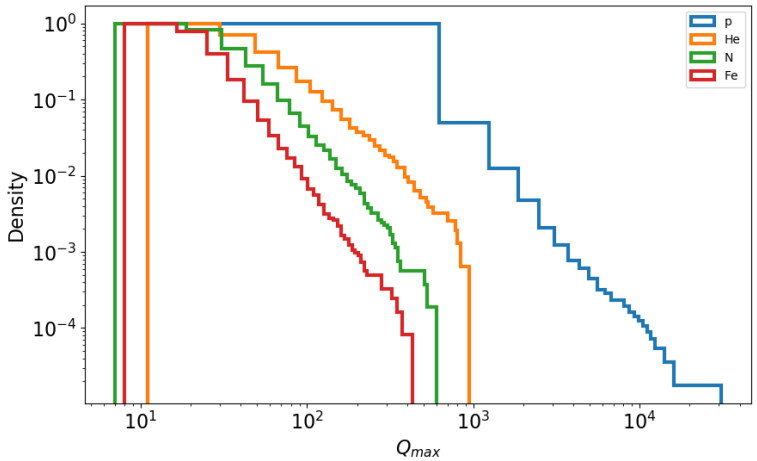


Figure 17: Distribution of Q_{max}

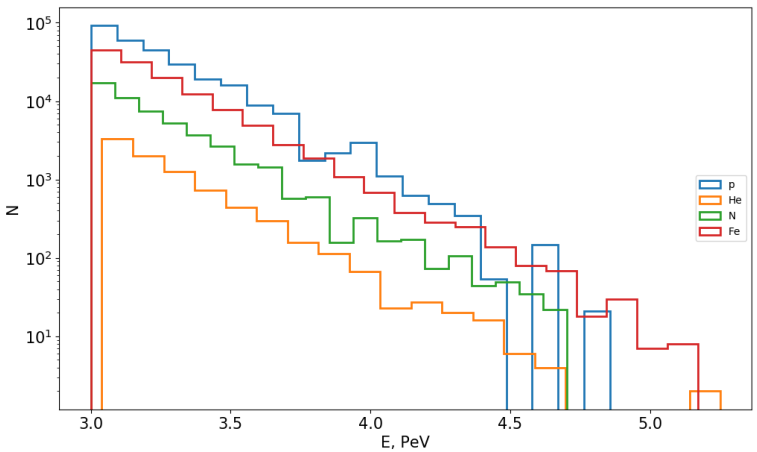


Figure 16: Distribution of E_0^{rec}.

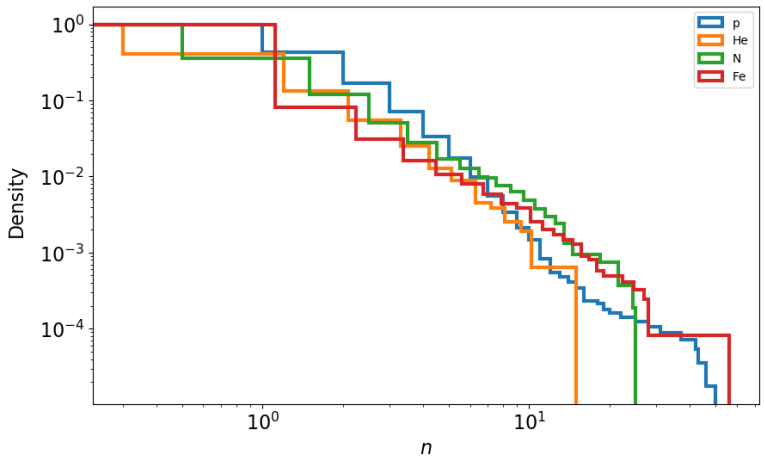


Figure 18: Distribution of N_n

Primary particle identification

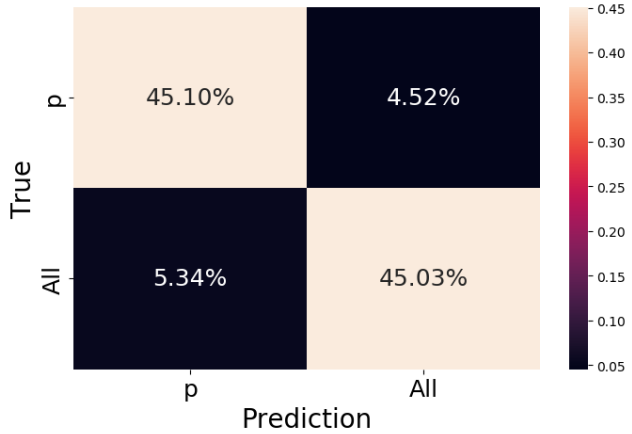


Figure 19: Confusion matrix over all validation set

Selection of events with low data and knowledge uncertainty.
About $\sim$ 48% of events were dropped.

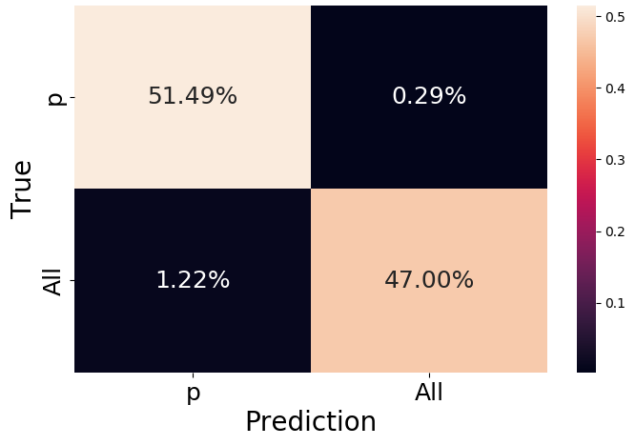


Figure 20: Confusion matrix of selected events

Metric	Before, %	After, %
Precision	90.1	**98.5**
Recall	90.1	**98.5**
F1-score	90.1	**98.5**

Conclusions

- ML algorithms allows to improve reconstruction performance of the EAS parameters.
- The presented algorithms allow us to estimate the uncertainty of the found parameters of the EAS and will be tested on different models of hadron interactions in the future.

Thanks for your attention!

Maximum likelihood estimation using some analytic LDF parametrization:

- $\rho_e(r,s) = \frac{1}{r_m^2}\frac{\Gamma(4.5-s)}{2\pi\Gamma(s)\Gamma(4.5-2s)}(\frac{r+\delta}{r_m})^{s-2}(1+\frac{r+\delta}{r_m})^{s-4.5}$ — NKG, where 0.5 $\leq s \leq$ 1.5;
- $\rho_e(r) = \frac{0.28}{R_{ms}^2}(\frac{r}{R_{ms}})^{-1.2}(1+(\frac{r}{10R_{ms}})^2)^{-0.6}(1+\frac{r}{R_{ms}})^{-3.33}$ — Scaling formalism[e];
- $\rho_e(r,s_\perp) = m^{-2}\rho_{NKG}(\frac{r}{m},s_\perp)$ — Uchaikin distribution, where m = 0.78 - 0.21 $s_\perp$ and 0.6 $\leq s_\perp \leq$ 1.8;
- $\rho_e(r,s) = \frac{1}{r_m^2}\frac{\Gamma(4.5-s)}{2\pi\Gamma(s)\Gamma(4.5-2s)}(\frac{r+\delta}{r_m})^{s+\alpha(r)-2}(1+\frac{r+\delta}{r_m})^{s+\alpha(r)-4.5}$ — NKG with local age[f], where 0.5 $\leq s \leq$ 1.5.

[e] Lagutin A. A. (2002) *Electron lateral distribution in air showers: scaling formalism and its implications.* Journal of Physics G: Nuclear and Particle Physics, 28(6), 1259.

[f] Capdevielle, J. N., Gawin, J. (1982). *The radial electron distribution in extensive air showers.* Journal of Physics G: Nuclear Physics, 8(9), 1317.

In order to select best LDF for array we simulate position of shower cores in circle with radius 15 m from center of array and select events that lie inside array borders. Next we estimate core resolution and metrics from ML that measure selection accuracy (precision) and size of selected events (recall).

TP - core selected inside array borders (inside array in model)
FN - core selected outside array borders (inside array in model)
FP - core selected inside array borders (outside array in model)

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision+Recall}$$

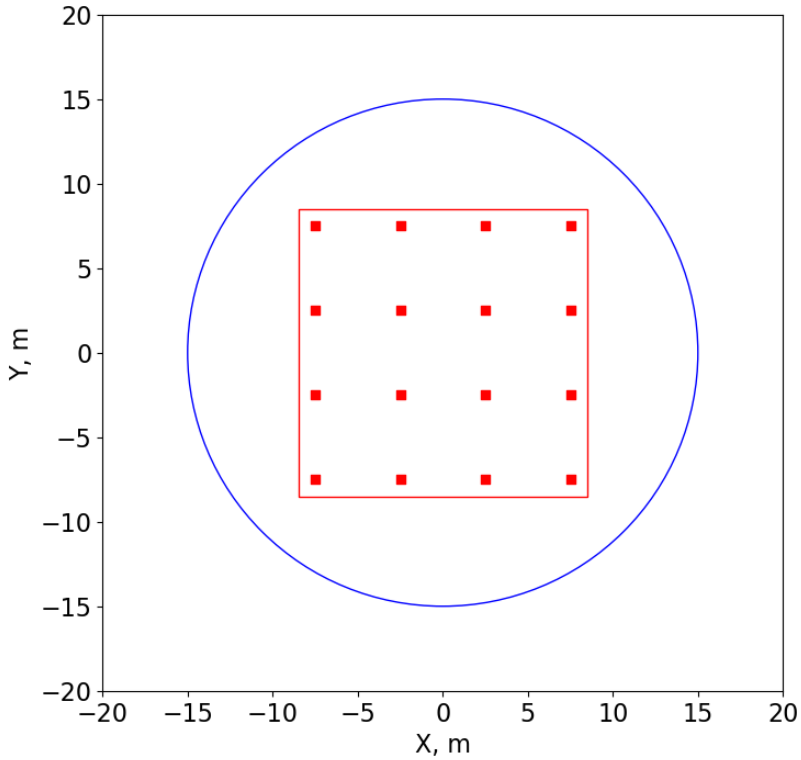


Figure 21: Selection area (red square) and simulation area (blue circle)

Shower core location error: $\Delta R = \sqrt{(X_{rec.} - X_{true})^2 + (Y_{rec.} - Y_{true})^2}$

Core resolution - quantile of 68% ΔR distribution. Comparison of metrics for presented LDFs:

Metric	NKG	Uchaikin	NKG s(r)
Core resolution	2.1 m	1.82 m	**1.7 m**
Precision	71%	72%	**73%**
Recall	81%	81%	**82%**
F1-score	75%	76%	**77%**

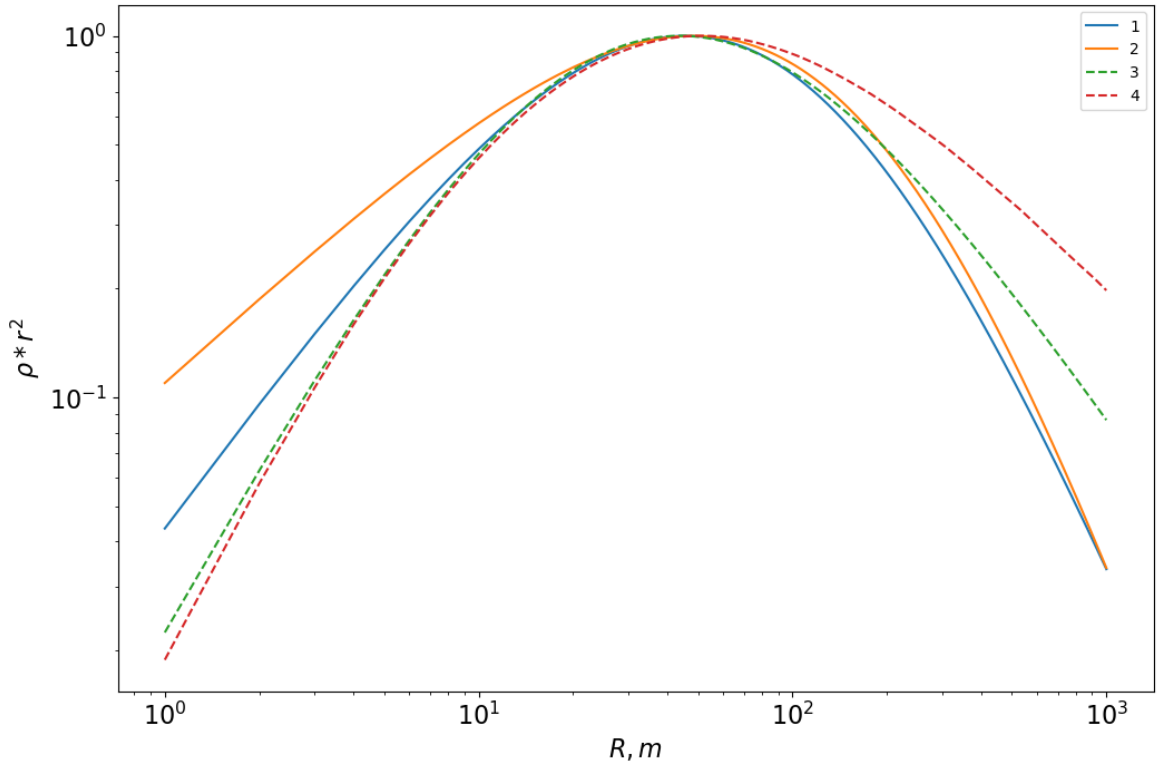


Figure 22: Normalized LDF's; (1) - NKG with s = 1.2, (2) - Lagutin scaling for $E_0 = 10^{15} eV$, (3) - Uchaikin distribution $s_\perp = 1.6$, (4) - NKG with local age (s = 1.2)

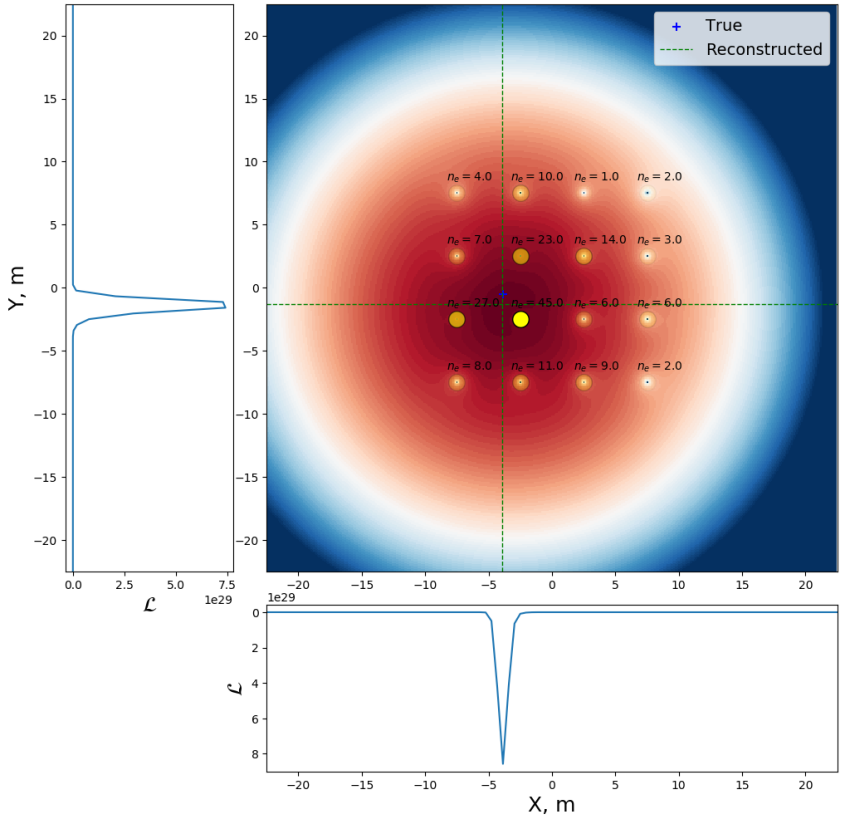


Figure 23: Event with a small ΔR error

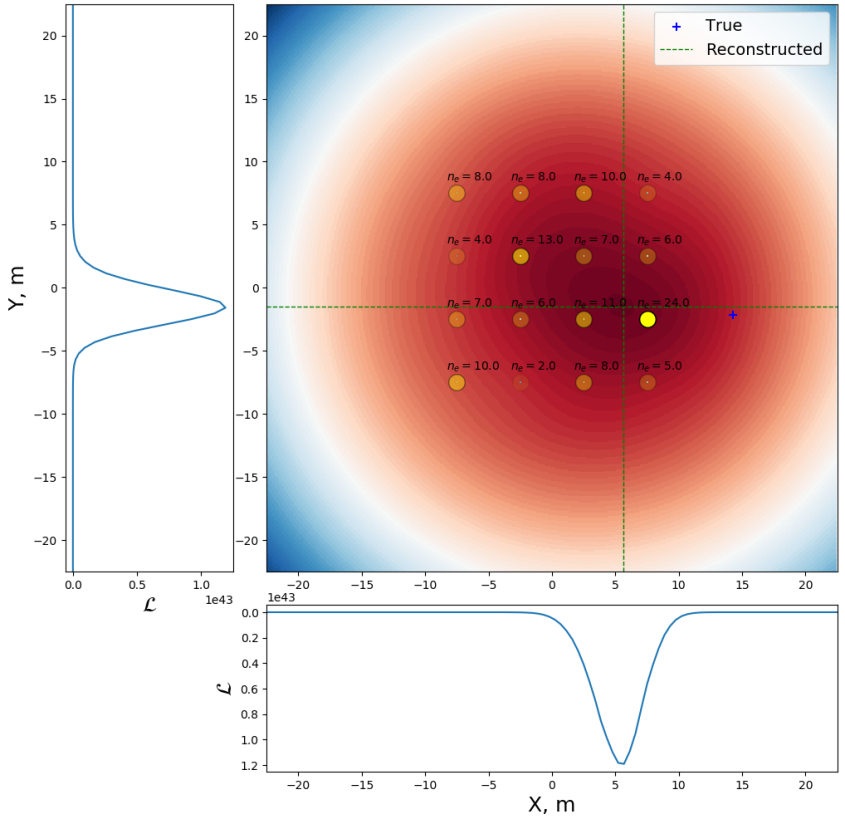


Figure 24: Event with a big ΔR error