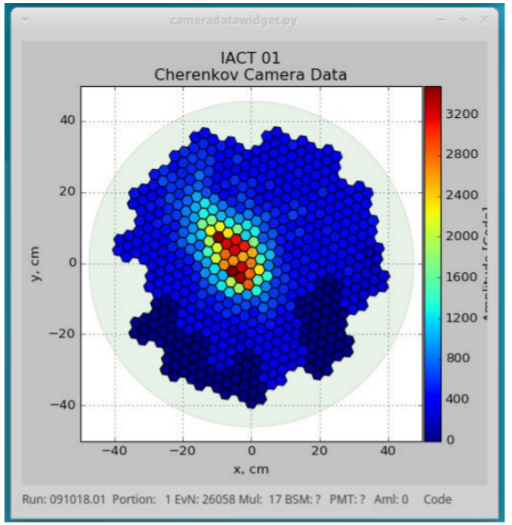
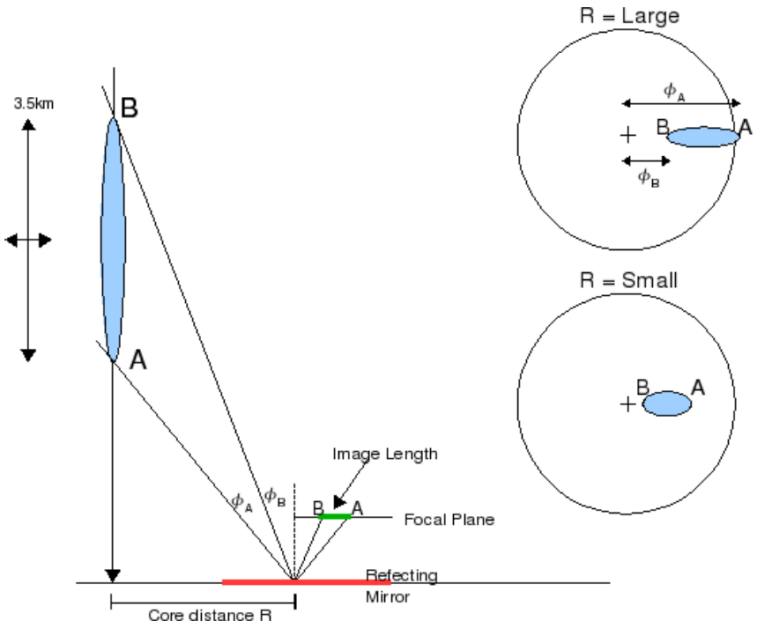
Gamma/hadron separation for a ground based IACT (imaging atmospheric Cherenkov telescope) in experiment TAIGA using machine learning methods Random Forest.

M. R. Vasyutina, L.G Sveshnikova
for TAIGA collaboration.

Image formation

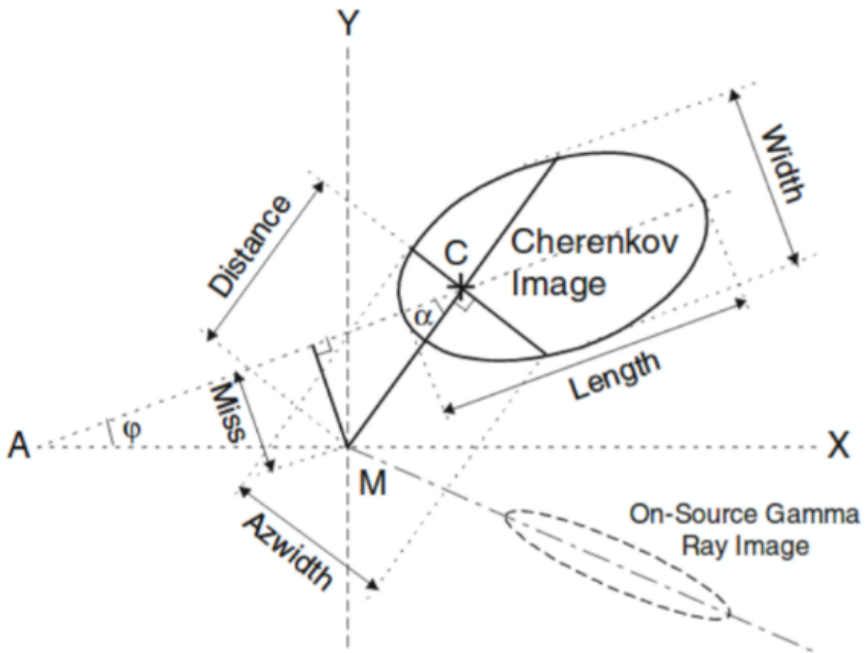




General view of the TAIGA-IACT telescope (a) and recorded images (b) of atmospheric showers by the telescope camera

Displaying a shower on the camera plane through a reflective mirror aimed at the source. Angle φ_B is the angle between the light from the top of the shower, point B, and the optical axis of the reflecting mirror, angle φ_A is the angle between the light from point A of the shower and the optical axis of the reflecting mirror. The camera image is shown for two R values.

Hillas parameters

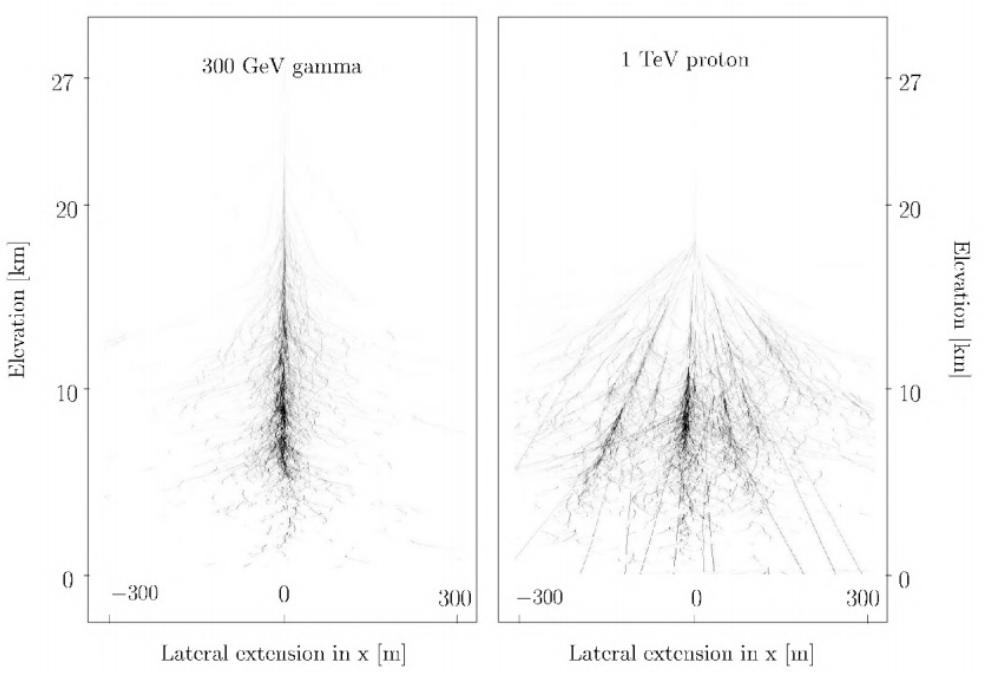


Schematic of an EAS elliptical image formed on the IACT matrix, where C is the center of gravity of the ellipse.

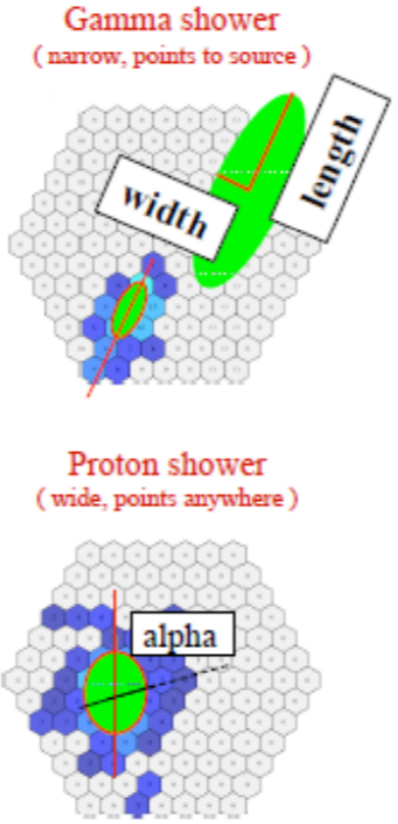
These parameters include:

- information about the shape of the ellipse (semi-major axis of the ellipse - length, semi-minor axis - width and azimuth width (Azwidth));
- location and orientation of the ellipse relative to the center of the camera (center of gravity of the ellipse, distance from the center of the camera (Distance), error (Miss), angle α and angle φ);
- Parameters characterizing the total number of photoelectrons in the image and their distribution over the ellipse (image size (Size), concentration (Concentration) and asymmetry coefficient);

Differences in Hill's parameters for different particles



Picture of the development of an EAS from a primary gamma quantum and a proton



Images from gamma quantum and proton

Decision tree visualization (max depth = 3)

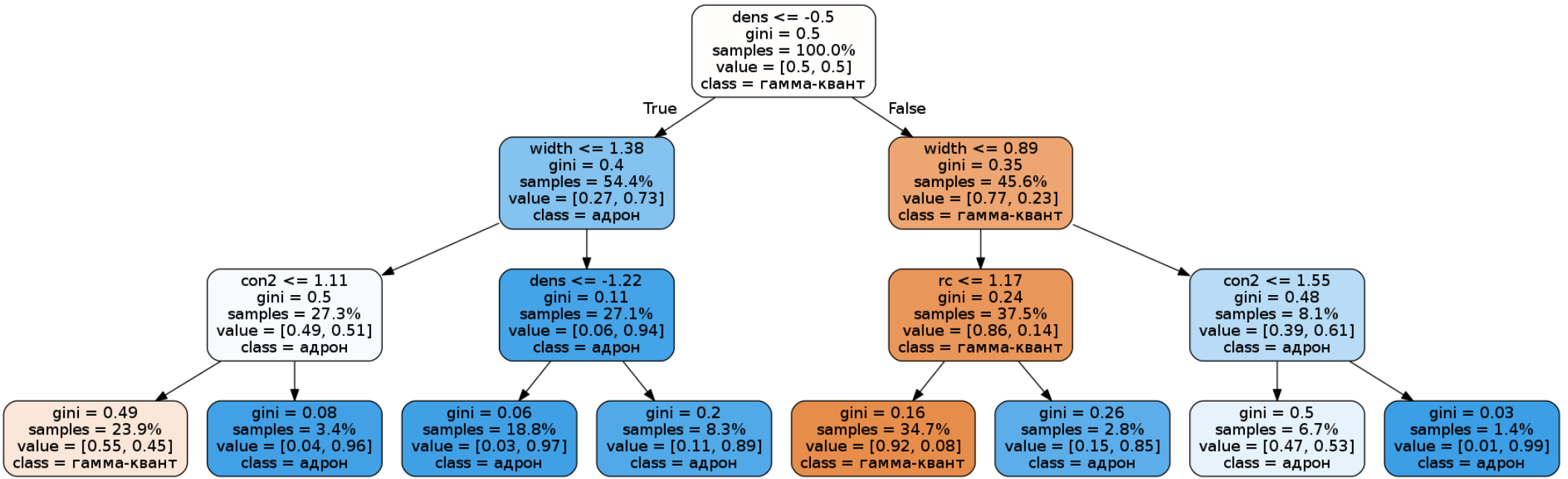


Image parameters

- Rc — distance to the weighted center of the image from the center of the camera with coordinates [0,0];
- Width — Hillas parameter;
- Con2 — concentration of light in pixels with maximum amplitude relative to the entire size;
- Dens — the ratio of the total number of photoelectrons in the image (Size) to the number of pixels in the image;

Gini criterion

$$I_G(n) = 1 - \sum_{i=1}^{J} (p_i)^2$$

The Gini contamination of node n is 1 minus the sum of the ratios of the class to the total number of samples pi squared for each of the set of classes J (in our case, 2 classes).

Possible ways to solve the problem of retraining

1) Limiting the maximum value of various tree parameters (decreasing variability, increasing error)
2) An increase in the number of trees (a decrease in the overall variability, without an increase in error)

$$l_{tree\,i,\,term.-node\,j} = \frac{N_h}{N_h + N_g} \qquad h = \frac{\sum_{i=1}^{n_{tree}} l_i}{n_{tree}} \qquad g = 1 - h$$

Random forest

- A Breiman random forest is an ensemble of decision trees, each of which is built on the basis of a bootstrap sample (a random sample of the same size as the original training sample, with return) from the original training sample (bugging), and only a fraction of the randomly selected signs. In addition, a complete tree is built (no truncation). The classification of trees in the ensemble is carried out by a majority vote.

Increased accuracy

- Using scaled parameters (rc, width, con2, dens);

$$\omega_{i,scaled} = \frac{\omega_i - \bar{\omega}_i}{\sigma_{\omega_i}}$$

Selection of optimal characteristics of trees and the value of the "gamma" parameter

Q-factor $$Q = \frac{\varepsilon_g}{\sqrt{\varepsilon_h}}$$

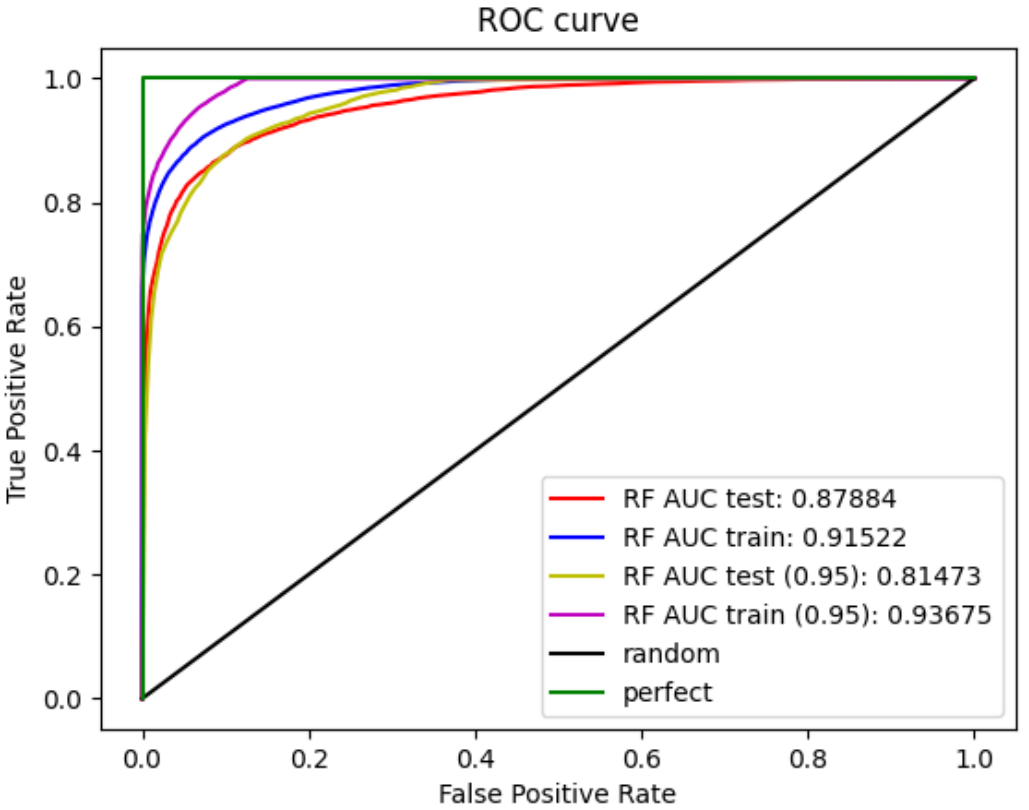
$$\varepsilon_g = \frac{N(g \to g)}{N(g \to g) + N(g \to h)} = \frac{N(g \to g)}{N_g(primary)}$$

$$\varepsilon_h = \frac{N(h \to g)}{N(h \to h) + N(h \to g)} = \frac{N(h \to g)}{N_h(primary)}$$

-suppresion efficiency

Roc - curve

The ratio between the share of objects from the total number of feature carriers, correctly classified as carrying a feature (true positive rate, TPR), and the share of objects from the total number of objects that do not carry a feature, erroneously classified as carrying a feature (false positive rate, FPR)

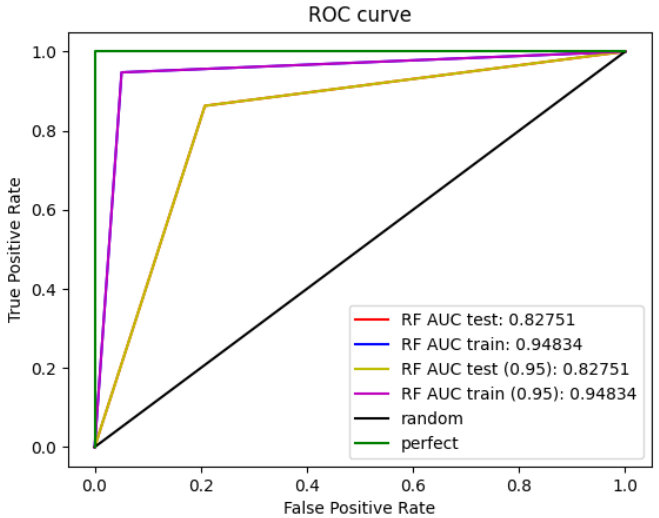


The training sample contains 18438 particles of each type, the test sample contains 4609 particles of each type

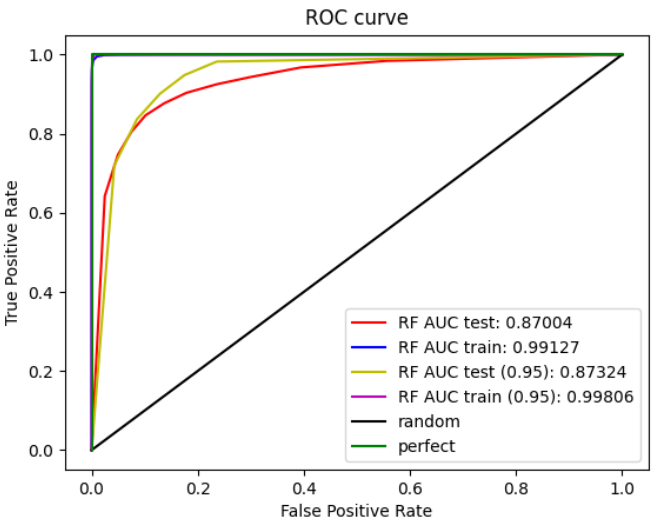
	g > 0.5						
N trees	g->g	h->h	h->g	g->h	ε_g	ε_h	Q factor
1	3651	3977	632	958	0.79215	0.13712	2.13919
10	3979	4041	568	630	0.86331	0.12324	2.45921
100	3926	4105	504	683	0.85181	0.10935	2.57592
1000	3951	4100	509	658	0.85724	0.11044	2.57955
10000	3951	4096	513	658	0.85724	0.11130	2.56948

	g > 0.95						
N trees	g->g	h->h	h->g	g->h	ε_g	ε_h	Q factor
1	3651	3977	632	958	0.79215	0.13712	2.13919
10	2042	4041	73	630	0.44305	0.01774	3.32598
100	1761	4105	40	683	0.38208	0.00965	3.88942
1000	1825	4100	46	658	0.39596	0.01110	3.75917
10000	1839	4096	46	658	0.39900	0.01111	3.78618

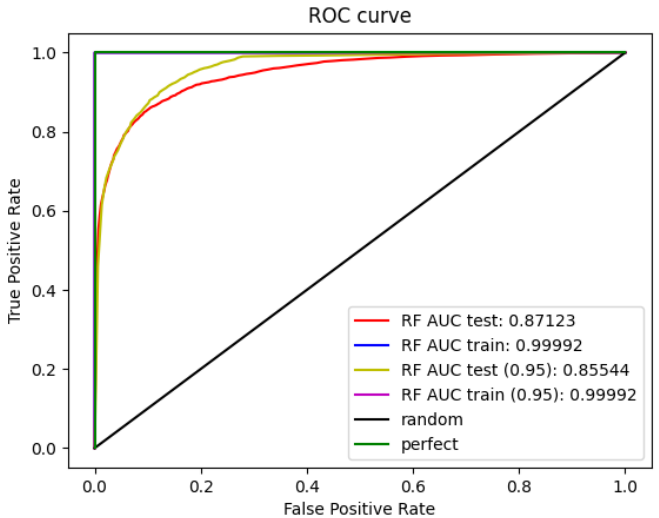
ROC-curve



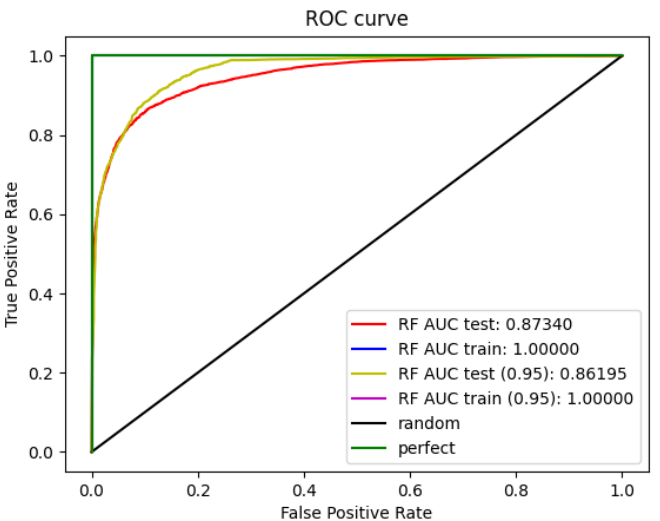
1 tree



10 trees



100 trees



1000 trees

Selection of the value of the limiting parameter "gamma"

	Number of trees = 100						
Threshold g	g->g	h->h	h->g	g->h	ε_g	ε_h	Q factor
0.8	3006	4105	180	683	0.65220	0.04201	3.18216
0.85	2744	4105	131	683	0.59536	0.03093	3.38548
0.9	2360	4105	82	683	0.51204	0.01958	3.65889
0.95	1761	4105	40	683	0.38208	0.00965	3.88942
0.98	1128	4105	21	683	0.24474	0.00509	3.43050

Selection of EAS image parameters

Size threshold

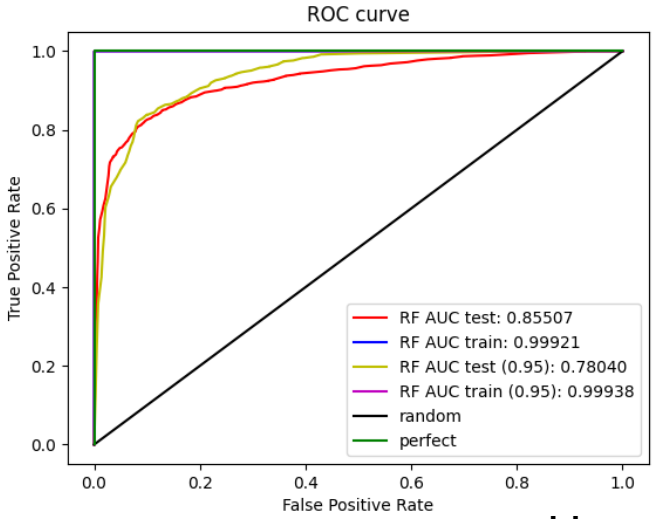
Size	train	test	g->g	h->h	h->g	g->h	ε_g	ε_h	Q factor
60	71572	17892	2505	7662	101	1836	0.28001	0.01301	2.45490
125	36876	9218	1761	4105	40	683	0.38208	0.00965	3.88942
140	31952	7988	1575	3578	33	537	0.39434	0.00914	4.12505
150	29148	7286	1480	3271	26	493	0.40626	0.00789	4.57483
160	26644	6660	1374	2991	30	443	0.41261	0.00993	4.14054
170	24410	6102	1281	2757	25	394	0.41986	0.00899	4.42910
180	22514	5628	1216	2550	22	358	0.43213	0.00855	4.67233

RC Threshold (Size> 125)

Rc	train	test	g->g	h->h	h->g	g->h	ε_g	ε_h	Q factor
23	36876	9218	1761	4105	40	683	0.38208	0.00965	3.88942
18	35702	8924	1526	3998	36	653	0.34200	0.00892	3.62028
13	32264	8064	1330	3636	23	520	0.32986	0.00629	4.16053
8	17666	4416	1016	2040	20	217	0.46014	0.00971	4.66996
5	3560	890	175	423	1	55	0.39326	0.00236	8.09769

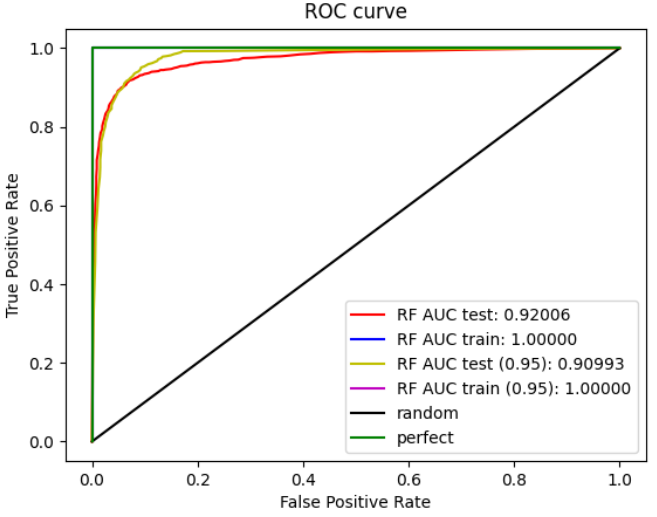
Add or remove some parameters

Param.	test	g->g	h->h	h->g	g->h	ε_g	ε_h	Q factor
-	4416	1016	2040	20	217	0.46014	0.00971	4.66996
- Dens		453	1911	16	343	0.20516	0.00830	2.25154
+ Size		960	2055	16	200	0.43478	0.00773	4.94655



Without Dens

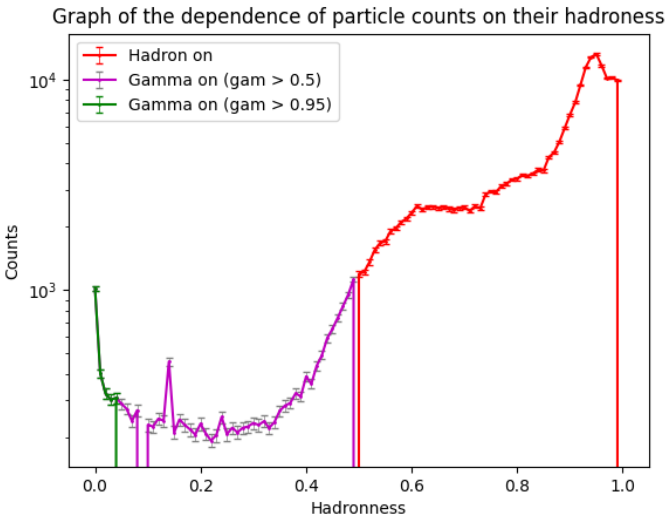
Here and below we use Size>125



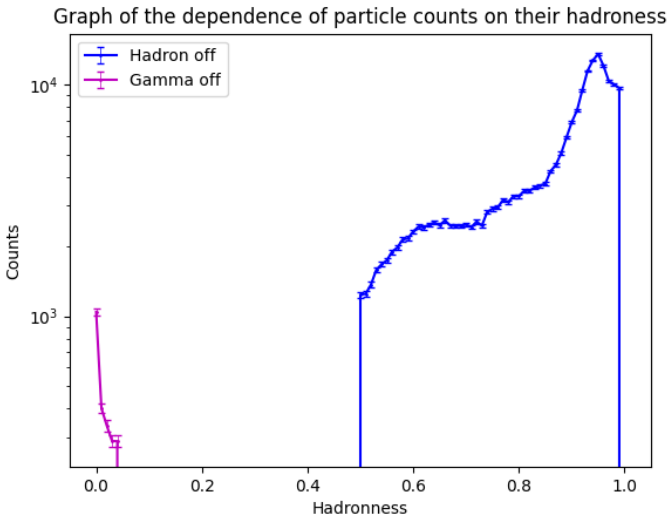
With Size

Working with experimental data

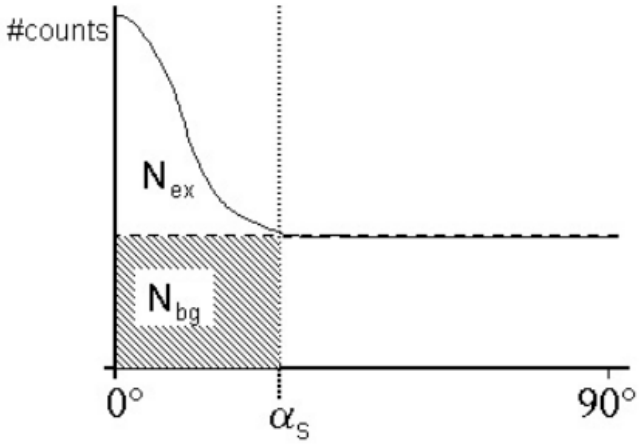
- When limited by Size> 125, Rc0 <18 (rc <8) of the total number of 666462 events, only 232087 are analyzed
- These results are considered for Size> 125

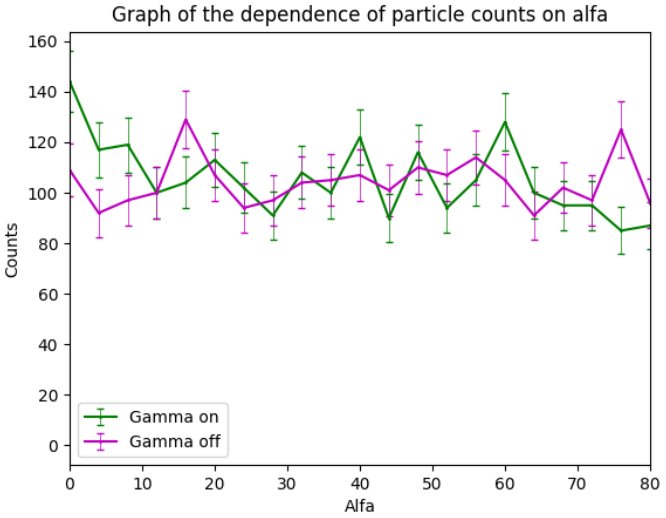


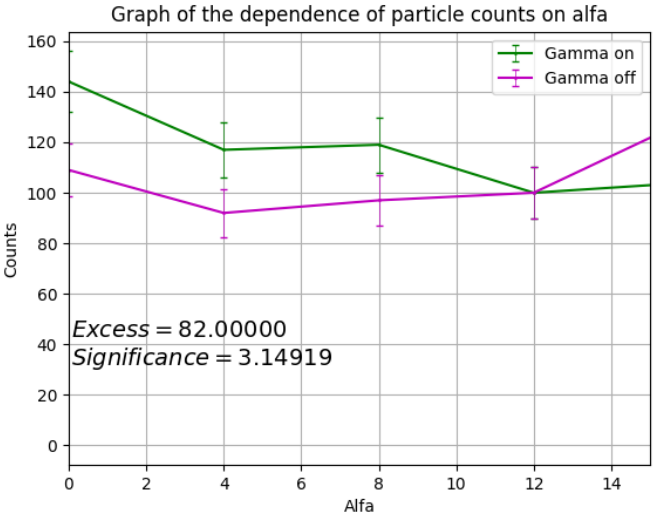
For ON data



For OFF data







These schematic plots show the calculation of N_{ex} and N_{bg}. The horizontal, dashed line, which represents the background level, can be assumed to be constant only in the simplest case.

Experimental result for the dependence of the amount of gamma quanta on alpha.

Experimental result for the dependence of the amount of gamma quanta on alpha. (In the region of alpha <15º)

$$Excess = N_{ON} - N_{OFF}$$

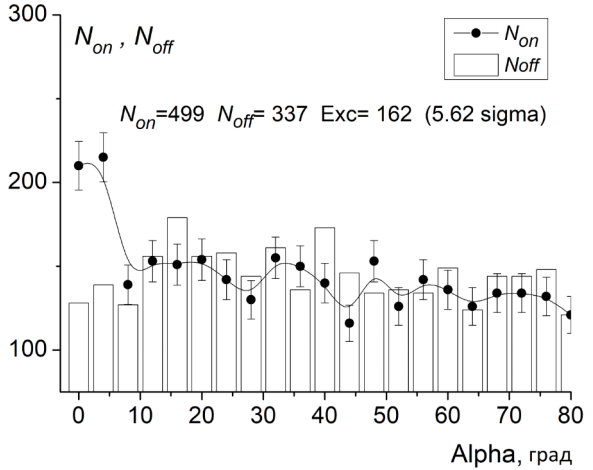
$$Significance = \frac{N_{ON} - N_{OFF}}{\sqrt{N_{ON} + N_{OFF}}}$$

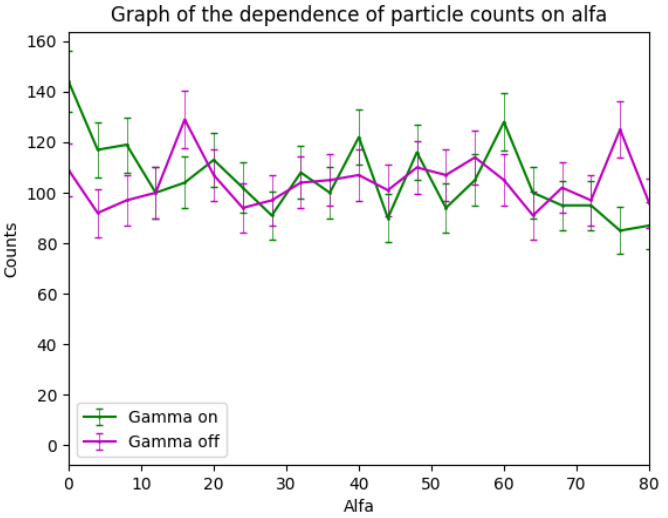
The results obtained with the same threshold for Size in the region of alfa <10º in the TAIGA experiment:

N_{ON}= 490, N_{OFF}= 337, the excess is Exc = 162 events with a significance of 5.62

The results of this work:

N_{ON}= 380, N_{OFF}= 298, the excess is Exc = 82 events with a significance of 3.15

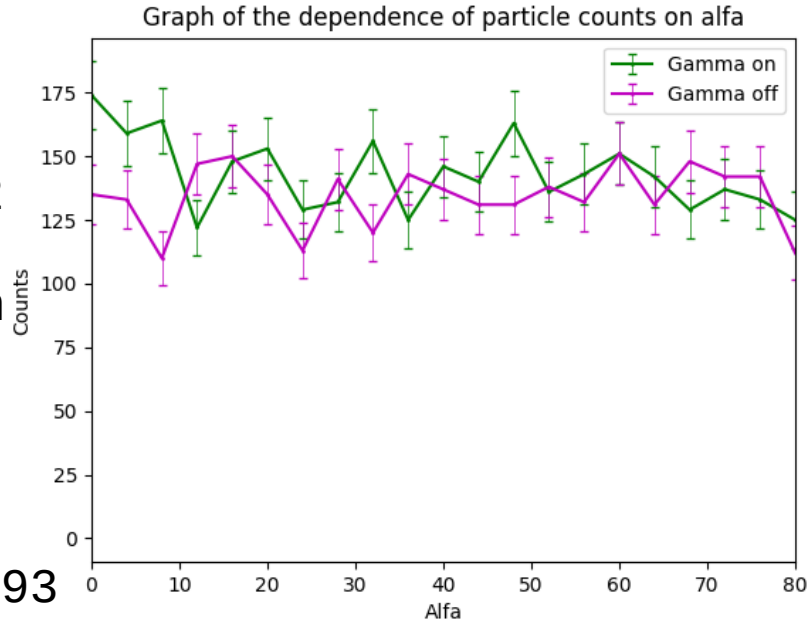




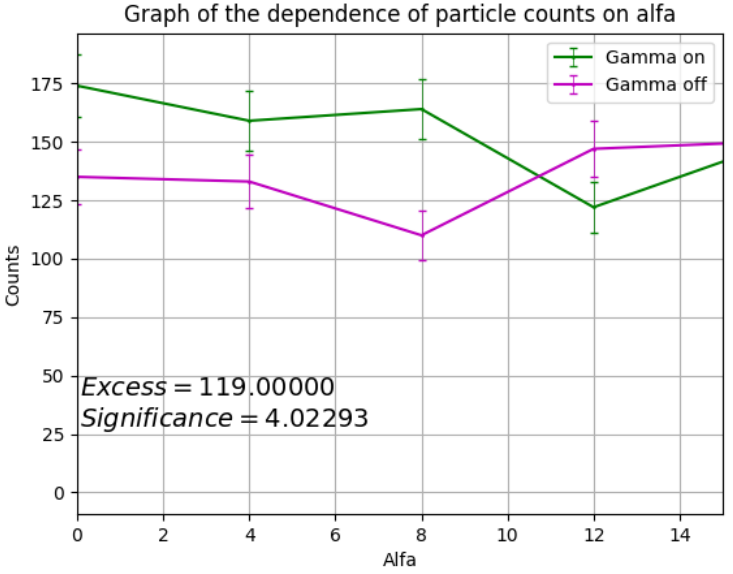
Using OFF sampling instead of MK hadrons for training

When testing:
ε_g =0.58967
ε_h = 0.01915
Q (0.95)=4.26162

When analyzing an experiment:
N_{ON} = 497;
N_{OFF} = 378;
Excess = 119
Significance = 4.02293



Experimental result for the dependence of the amount of gamma quanta on alpha.

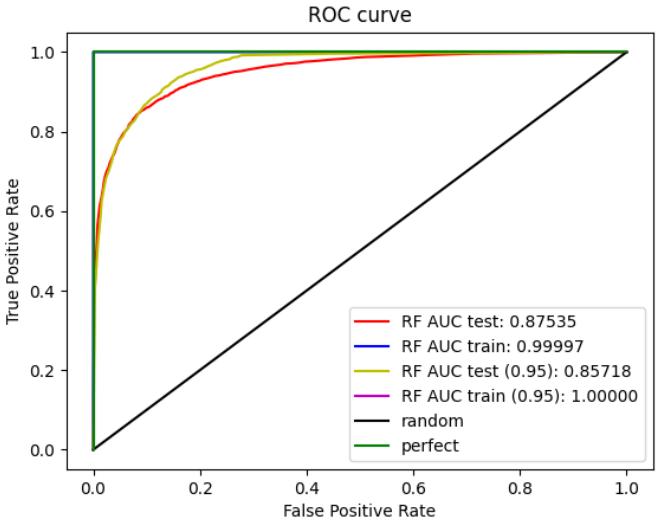


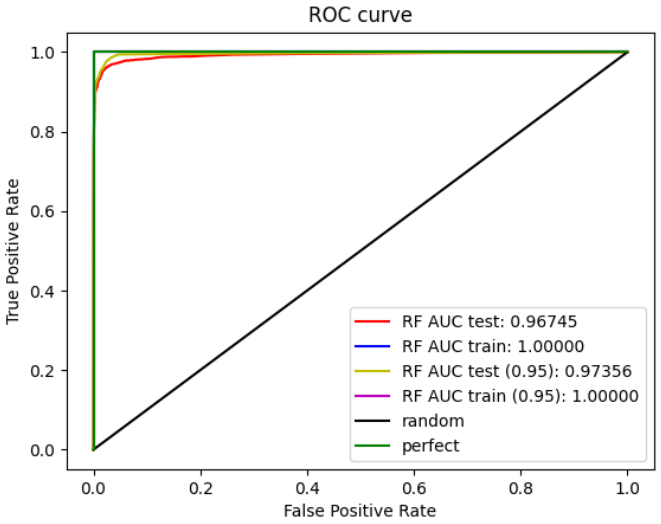
Experimental result for the dependence of the amount of gamma quanta on alpha. (In the region of alpha <15º)

Adding the alfa parameter

Param.	g->g	h->h	h->g	g->h	ε_g	ε_h	Q factor
Without alfa	960	2055	16	200	0.43478	0.00773	4.94655
With alfa	1530	2117	15	79	0.69293	0.00704	8.26115

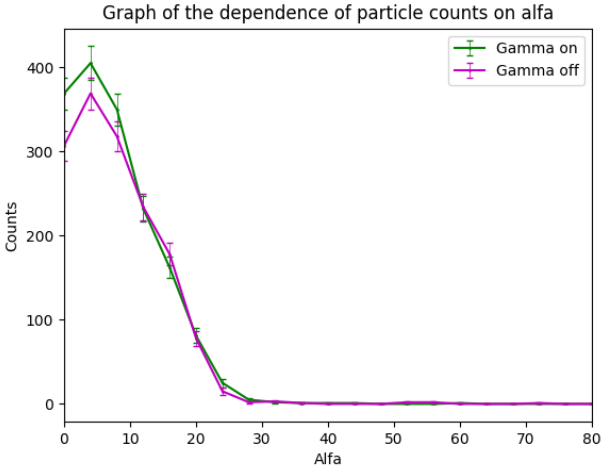
ROC curves before (left) and after (right) adding the alfa parameter.



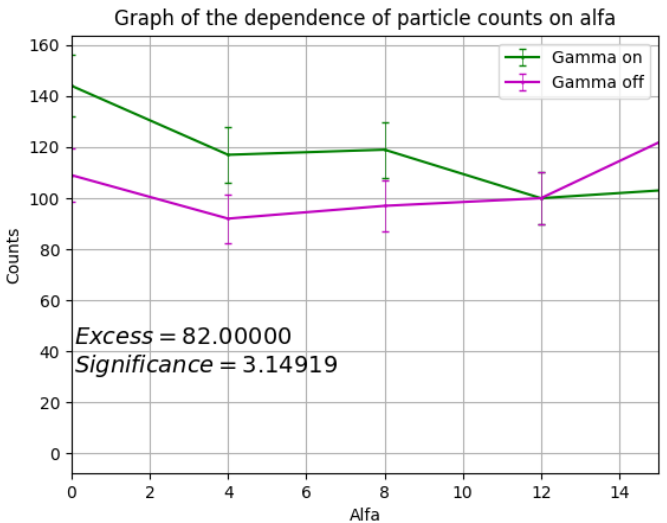


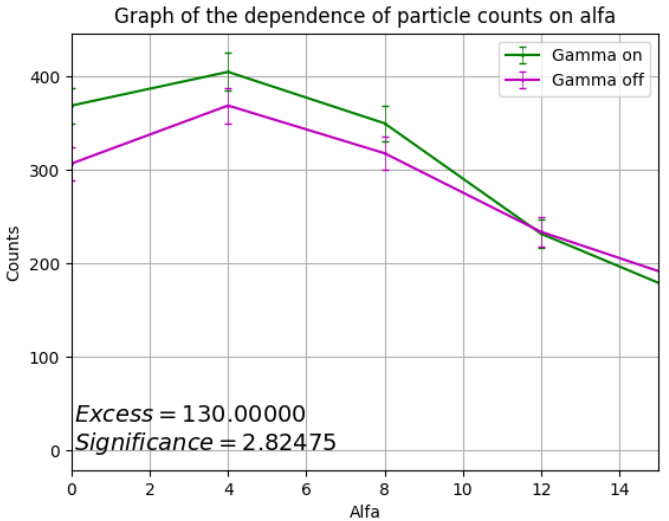
Adding the alfa parameter

Param.	N_{ON}	N_{OFF}	Excess	Significance
Without alfa	380	298	82	3.14919
With alfa	1124	994	130	2.82475



Dependences of the number of particles from alfa before (left) and after (right) adding the alfa parameter for alfa <15 (bottom) and not limited alfa (top).





conclusions

- In this work, a program has been mastered, which includes the RANDOM FOREST machine learning algorithm, and is adapted to perform the task of suppressing the hadronic background in the TAIGA experiment.
- On the samples obtained from the Monte Carlo data for primary particles, the optimal settings of the program for the given problem, such as the number of trees, etc., are investigated and the parameters of the images are selected. It is shown that the method produces stable results and is robust to input parameters. The classification accuracy expected from the input samples is obtained.
- A similar optimization was carried out on the experimental samples ON (when the telescope is directed to the source) and OFF (when the telescope is directed to the background of the sky). The dependence on the alpha parameter showed the possibility of separating events initiated by gamma quanta from the background of hadrons.
- The obtained result is compared with the methods of background suppression by semi-empirical selection of parameters used in the TAIGA experiment to date. Compared to the empirical approach, the method adapts to the available multidimensional space with minimal human intervention: setting the number of trees, choosing a gamma limitation, etc. This simpler control and tuning of the gamma-ray shower extraction procedure is considered a general advantage of the RF method.
- The possibility of using samples for training and testing the Random Forest algorithm is considered, consisting of events from gamma quanta obtained from Monte Carlo (MC) modeling, and events from hadrons obtained from an experimental sample when the telescope is aimed at the background of the sky or at "anti-source" (OFF sampling).

Thanks for attention