Machine learning aided noise filtration and signal classification for CREDO experiment

Łukasz Bibrzycki, Olaf Bar, Marcin Piekarczyk,
Michał Niedźwiecki, Krzysztof Rzecki, Sławomir Stuglik
on behalf of CREDO collaboration

12-23 July 2021

Executive summary

The amount of candidate hits in the Cosmic Ray Extremely Distributed Observatory (CREDO) database currently surpasses 10^7. In order to sustain the users' commitment CREDO makes extensive use of gamification techniques like periodical Particle Hunters Competitions. These result in a considerable increase of data into the CREDO database, however, with the adverse effect of the surge of artefacts, ie. images related to hardware malfunction or simply to users' cheating. Therefore, efficient means of rejecting the non-cosmic-ray noise and identification of signals attributable to extensive air showers are necessary. To address these problems we discuss a Convolutional Neural Network-based method of artefact rejection and complementary method of particle identification based on common statistical classifiers as well as their ensemble extensions. These approaches are based on supervised learning, so we need to provide a representative subset of the CREDO dataset for training and validation. According to this approach over 2300 images were chosen and manually labeled by 5 annotators. The images were split into spot, track, worm (collectively named signals), and artefact classes. Then the preprocessing consisting of luminance summation of RGB channels (grayscaling) and background removal by adaptive thresholding was performed. For purposes of artefact rejection the binary CNN-based classifier was proposed which is able to distinguish between artefacts and signals. The classifier was fed with input data in the form of Daubechies wavelet transformed images. In the case of cosmic ray signal classification, the well-known feature-based classifiers were considered. As feature descriptors, we used Zernike moments of 8^{th} degree consisting of 25 components with additional feature related to total image luminance. For the problem of artefact rejection, we obtained an accuracy of 99%. For the 4-class signal classification, the best performing non-ensemble classifier, ie. $\nu-$Support Vector Classifier achieved a recognition rate of 98% for Spots, 95% for Tracks, 70% for Worms, and 90% for Artefacts.