Machine-Learning-Based Data Cleaning for Anomaly Searches in the CREDO Smartphone Detector Network

B. Izydorczyk
Henryk Niewodniczański Institute of Nuclear Physics, Polish Academy of Sciences

Motivation and Goal

One of CREDO's goals is to search for anomalies and large-scale correlations in secondary cosmic-ray detections recorded by a distributed network of smartphones. Such searches require a clean dataset, since even a relatively small number of false detections can produce artificial excesses and obscure potentially interesting events.
Smartphone-based measurements are affected by detector imperfections, light leaks and incorrect camera shielding, which can generate image artifacts resembling genuine particle detections. Although deterministic filters remove many simple artifacts, some remaining cases are difficult to describe using explicit rules.
Goal: improve data cleaning by combining deterministic filters with CNN-based image classification, reducing the number of false detections while preserving genuine particle candidates for subsequent anomaly and correlation searches.

CREDO and Data

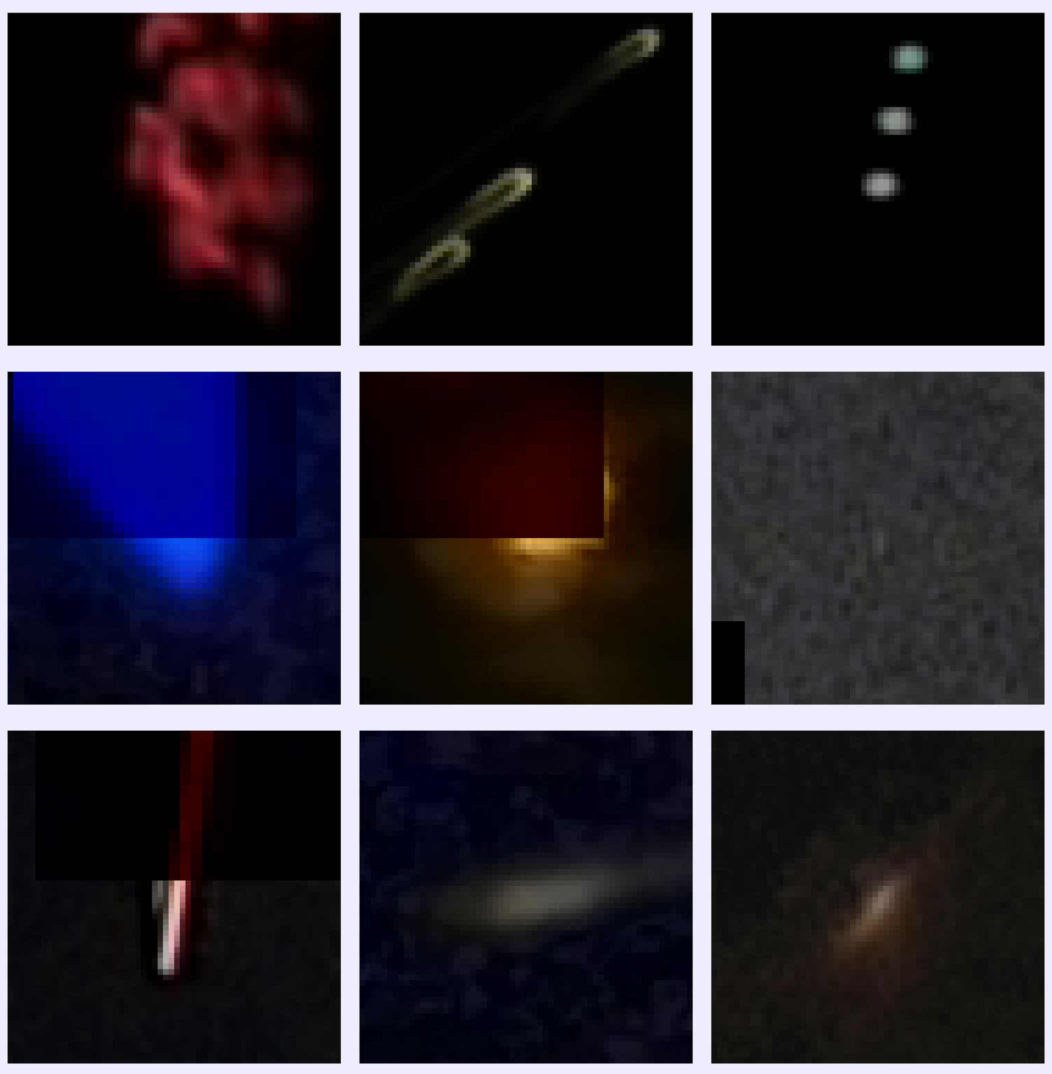
The Cosmic-Ray Extremely Distributed Observatory (CREDO) is a scientific project focused on fundamental research, including particle interactions and the structure of the Universe. As part of CREDO, smartphones and citizen participation are used to complement traditional methods of cosmic-ray detection.
The CREDO Detector app turns a smartphone into a small particle detector. Charged particles crossing the camera's CMOS sensor can produce characteristic bright tracks. Each recorded event contains information about the detection time and detector, together with a 64×64 image of the candidate track.
These images provide additional information about individual detections and can therefore be used to distinguish particle candidates from detector- or camera-related artifacts.

Examples of genuine particle candidates

Problems with Data

Smartphone cameras are not dedicated particle detectors and the recorded images contain a wide variety of non-particle signals. These may originate from light leaks, sensor defects, improper camera shielding or other camera-related effects.
Importantly, artifacts do not form a single, well-defined class. Their shapes, brightness distributions and spatial patterns can differ substantially, as illustrated below.

Examples of image artifacts



Simple artifacts can be efficiently removed using deterministic cuts based on quantities such as brightness, position or repeated sensor patterns. The difficulty arises for more complex cases, whose appearance may still be obviously non-particle-like to a human observer but cannot be captured by a small set of simple conditions.

Solution to the Problem

Instead of adding increasingly complex handcrafted rules, a CNN was used as a second filtering stage:

Rule-based filters → CNN classifier → cleaned dataset

Training set

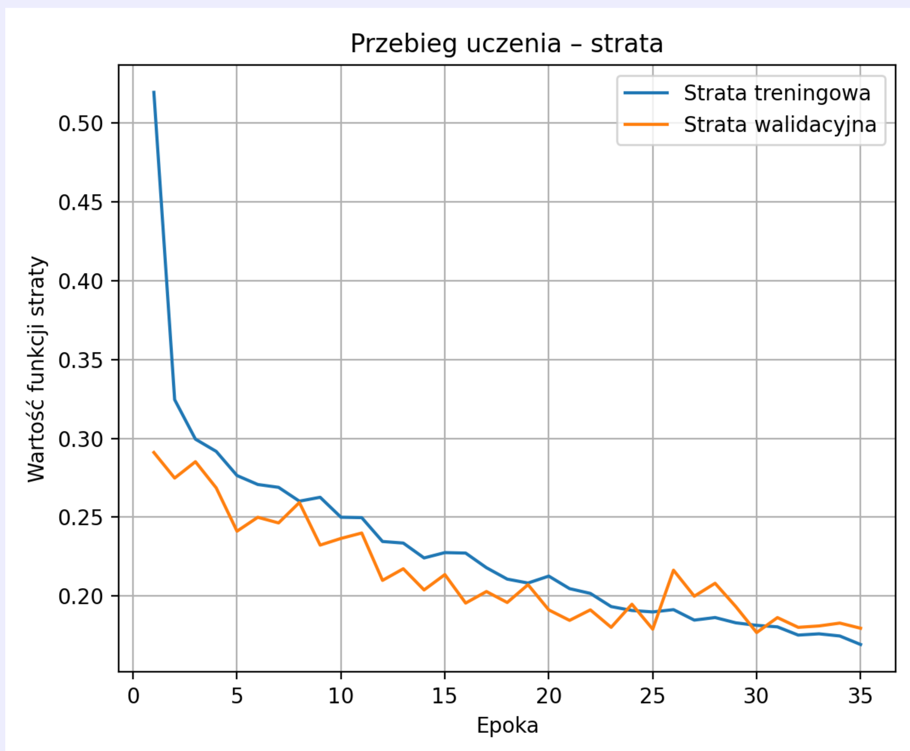
- manually labeled clear examples of genuine detections and artifacts;
- **10 000 images**: ∼4800 genuine and ∼5200 artifacts;
- ambiguous detections excluded;
- **80% / 20%** training–validation split;
- 64×64 grayscale images with random flips.

Model

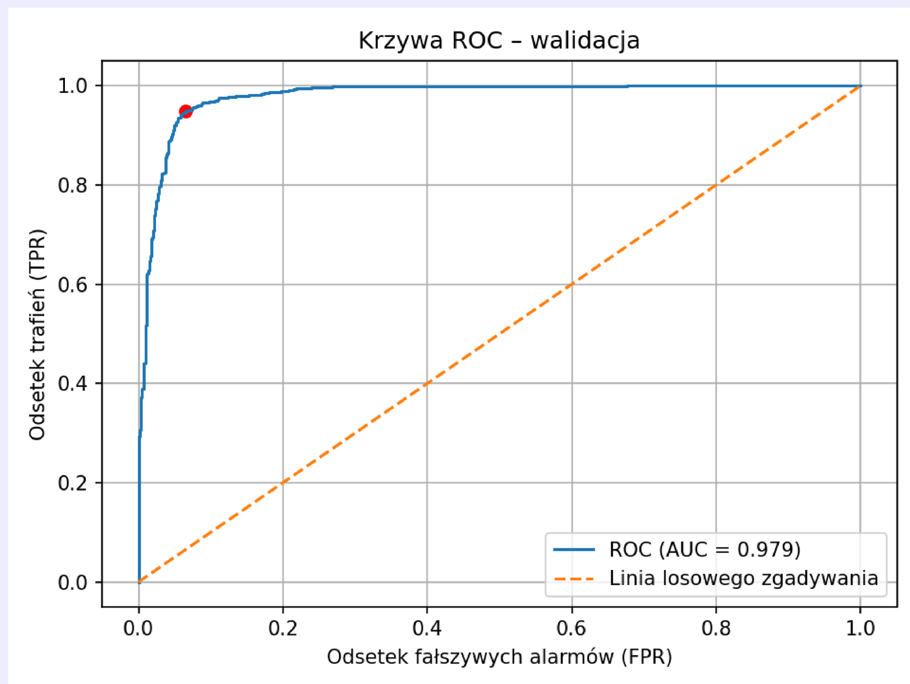
A simple CNN was sufficient: **3 convolutional layers** with 16, 32 and 64 filters followed by a dense classification layer.

Performance and Gain

The best-performing model was selected after **30 epochs**. The similar training and validation performance indicates good generalization to previously unseen detections.
The decision threshold is marked by the red dot on the ROC curve. Choosing this threshold allows the model to reject most artifacts while discarding only a small fraction of genuine detections.



Training and validation accuracy and loss. Further training leads to overfitting.



ROC curve for the validation dataset.

95%	93%	0.979	93.66%
Genuine detections retained	Artifacts rejected	ROC AUC	Validation accuracy

Decision threshold: $p_{\mathrm{cut}} = 0.60$

Main result: the CNN removes complex artifacts that are difficult to describe using deterministic rules while preserving most genuine particle candidates.

Conclusions and Outlook

Conclusions

- CNN complements rule-based filters by identifying artifacts that are visually recognizable but difficult to describe using explicit deterministic rules.
- The model generalizes well to previously unseen detections, with similar performance on the training and validation datasets.
- At the selected decision threshold, the CNN rejects most artifacts while preserving the vast majority of genuine particle candidates.

Outlook

- Apply the method to newer and larger CREDO datasets.
- Expand the labeled training set with detections from a wider range of smartphone models and operating conditions.
- Compare the current CNN with more advanced models and quantify the improvement over rule-based filtering alone.

Refernces

[1] Ł. Bibrzycki et al., *Symmetry* 12, 1802 (2020);
[2] Homola, P. et al. Cosmic-Ray Extremely Distributed Observatory. Symmetry 2020, 12, 1835.

Acknowledgements: The author thanks Prof. Aleksander Filip Żarnecki for supervision and the CREDO Collaboration for providing the experimental data.

LaTeX TikZposter