Using machine learning for data QA, PID, and fast simulations in ALICE

Łukasz Graczykowski, Tomasz Trzciński, Kamil Deja, Michal Glinka
for the ALICE Collaboration

Faculty of Physics
Warsaw University of Technology

DATA QUALITY ASSURANCE

- **Quality Assurance (QA)** and **Data Quality Monitoring (DQM)**: recording data of the highest quality and data taking procedure occurs as expected
- Currently, data quality assurance in ALICE TPC partially relies on **manual labour** of **highly qualified detector experts**
- **Label assignment** based on a set of features related to the statistical parameters of a run

Classification methods

Evaluated **machine learning** methods for quality label assessment:

- **Random Forest + ADABoost** - Adaptive ensemble classifier based on decision trees, predicting the class which was chosen with most of components.
- Support Vector Machine - non linear classifier, which tries to maximize the possible distance from decision boundary to any data samples.
- Naive Bayes Classifier - basic classification algorithm deriving straight from Bayesian theorem.
- Probabilistic neural networks - a Rosenblatt's perceptron which is trained to predict the probability of given sample belonging to each class.

Dataset

- **Over 1400 runs**
- Recorded by ALICE between January 2016 and July 2017
- 250 numerical attributes – statistical descriptions from trending.root
- **Ground truth**: samples manually annotated by experts with one of out of five labels:

Label	Nr of runs
Fully operational	795
Not set*	99
Off	373
Partial acceptance	110
Unusable	68

*Class Not Set is assigned when experts are not sure about the quality of a certain sample

Problem

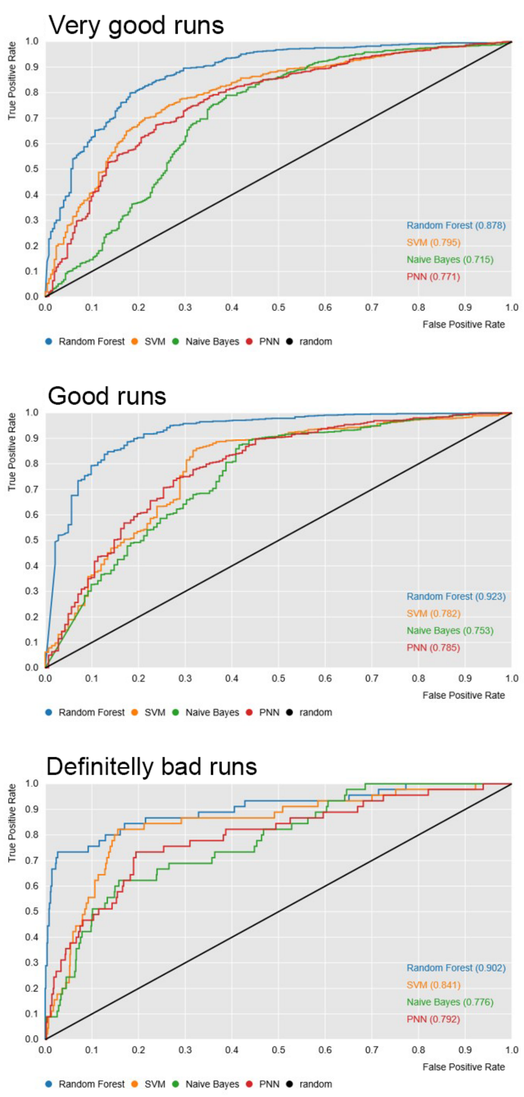
- Data split to **3 binary classification problems** corresponding to real-life use cases
- The goal: reduction of a **false positive rate**

Class	Definitely good	Good	Definitely bad
Fully operational	very good	good	To check
Partial Acceptance	To check	good	To check
Not set	To check	To check	To check
Unusable	To check	To check	bad

Experimental protocol

- data preprocessing: outliers removal, PCA dimensionality reduction
- automatic hyperparameter optimization
- adaptive boosting
- k-fold cross validation

Results



Conclusions

- **A Random Forest with adaptive boosting** significantly outperforms other methods across all use cases
- It allows to assign the quality label in **75% of the cases with over 95% accuracy**

FAST SIMULATIONS

- **Generative (GAN) models** used for **fast simulation**
- Widely used for generation of photo-quality artificial images
- The first step towards **semi-real-time** Quality Assurance solution for anomaly detection
- **Fast inference** of machine learning methods allows for fast generation of **plausible "healthy" detector** states
- Comparing real detector outputs with synthetic "healthy" indicates determines **anomalies**

Evaluated generative models

- Generative Aversarial Networks (GAN)
  - Multi Layered Perceptron (MLP) GAN
  - Long Short Term Memory(LSTM) GAN
  - Deep Convolutional GAN (DCGAN)
- Variational Autoencoder (VAE)
  - Standard Variational Autoencoder
  - Convolutional Autoencoder

To reduce visible aliasing, we implement a **progressive training** method for **DCGAN** model.

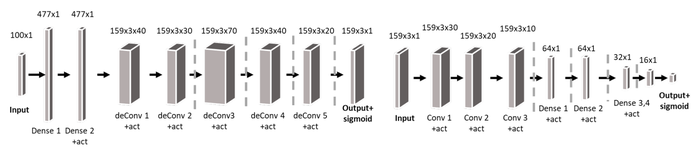


Figure 1: Architecture of the network for progressive training

Exemplar results

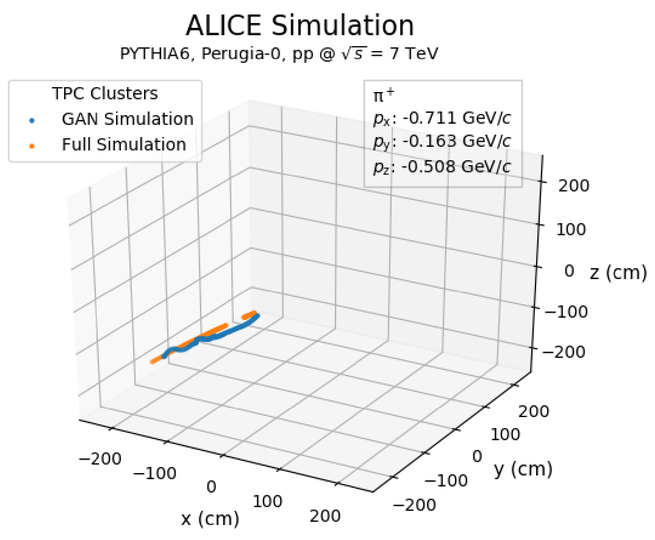


Figure 2: Pion TPC clusters simulated with DCGAN (blue) and full detector response with GEANT3 (orange).

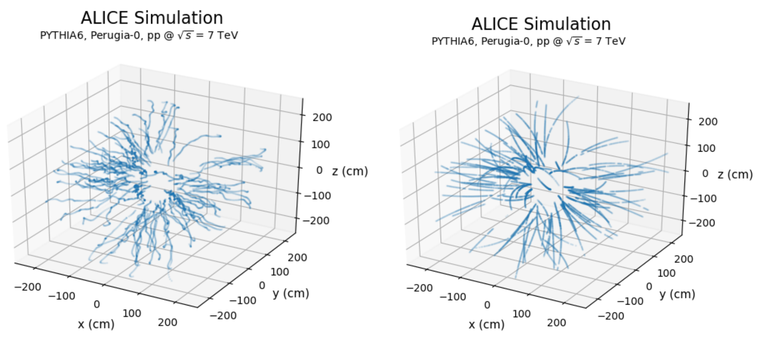


Figure 3: Simulation of TPC clusters with (left) conditional DCGAN method and (right) full detector response with GEANT3.

Quantitative comparison

Method	MSE (mm)	Speed up
GEANT3	0.085	1
VAE	37.415	10^4
cVAE	13.33	10
MLP	55.385	10^4
LSTM	54.395	10^4
DCGAN	26.18	10^2
proGAN	0.88	30

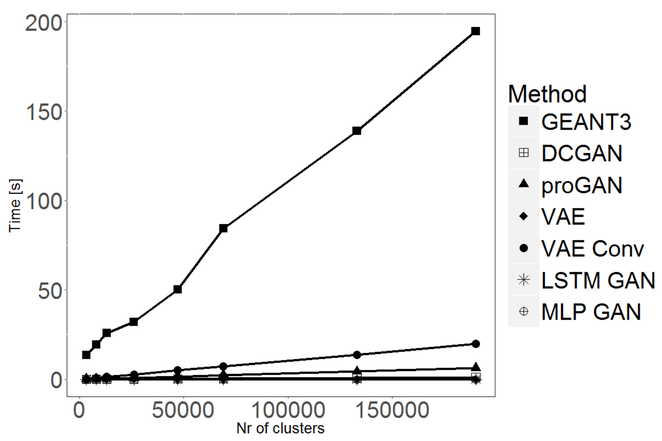


Figure 4: Performance of GAN methods w.r.t. traditional simulation.

Conclusions

- **Proof of concept** work showing promising results
- **Significant speed-up** w.r.t. traditional detector simulation
- Needs more tuning to get physical results
- Close cooperation with the CERN group developing GANs for EM calorimeters
- **Mutual benefits for the HEP and ML fields**

PARTICLE IDENTIFICATION

- Efficient identification of particles is crucial for many physics analyses
- Traditional PID methods are based on sub-omptimal "cuts"
- "Cuts" optimalisation is a time-consuming manual task
- Classification of particles is a perfect task for machine learning

Dataset

- PYTHIA6 Perugia-0, pp @ $\sqrt{s} = 7$ TeV with full detector response
- Selection of charged hadrons – pions, kaons, protons
- Signals from the TPC and TOF detectors used

Classification method

Traditional PID:

- $0.2 < p_T < 0.5$ GeV/c – TPC only, $N_{\sigma,\mathrm{TPC}} < 2$
- $p_T > 0.5$ – TPC and TOF, $\sqrt{{N^K_{\sigma,\mathrm{TPC}}}^2 + {N^K_{\sigma,\mathrm{TOF}}}^2} < 2$

Machine learning-based PID:

- Random Forest classifier

Exemplar results

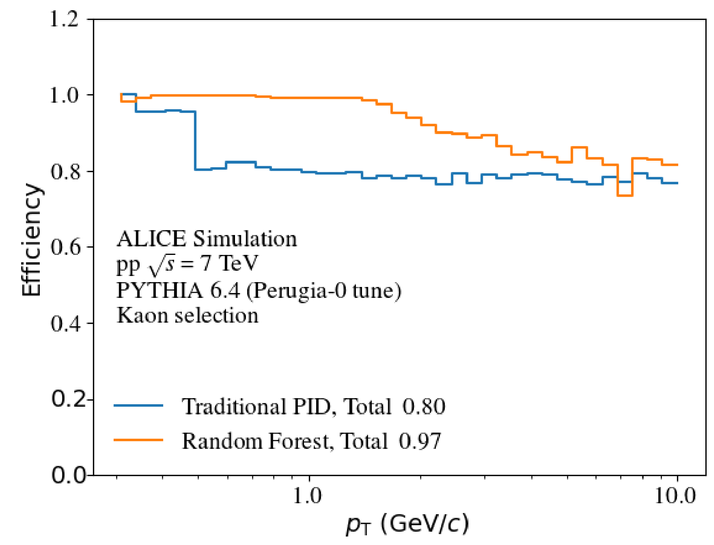


Figure 5: Efficiency of **charged kaon** selection by traditional and ML-based PID.

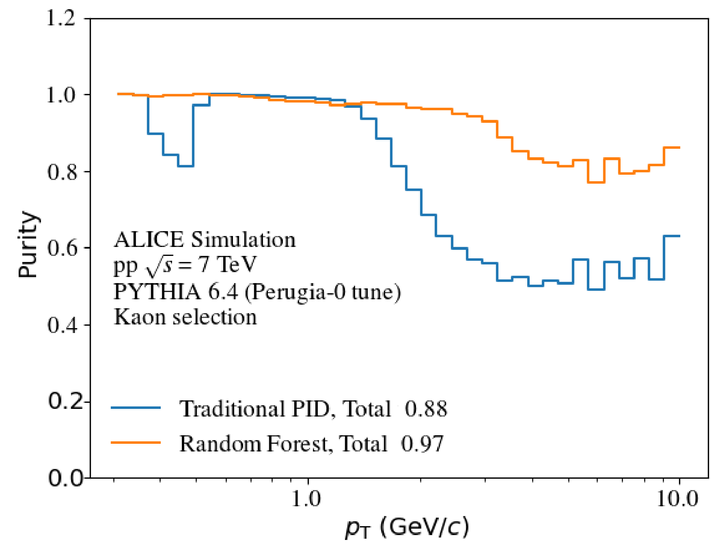


Figure 6: Purity of **charged kaon** selection by traditional and ML-based PID.

ML PID selection details

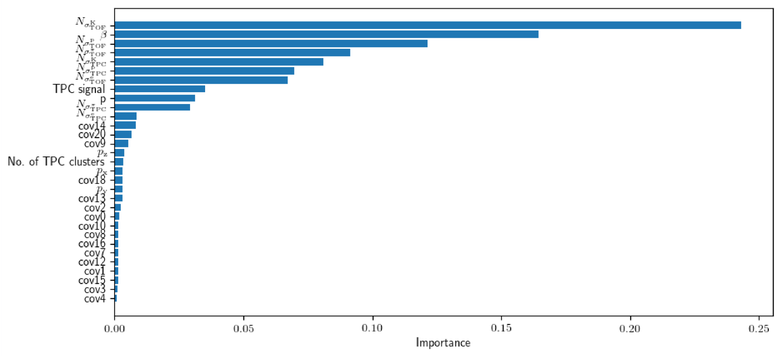


Figure 7: Importance of track parameters as inputs for Random Forest decision.

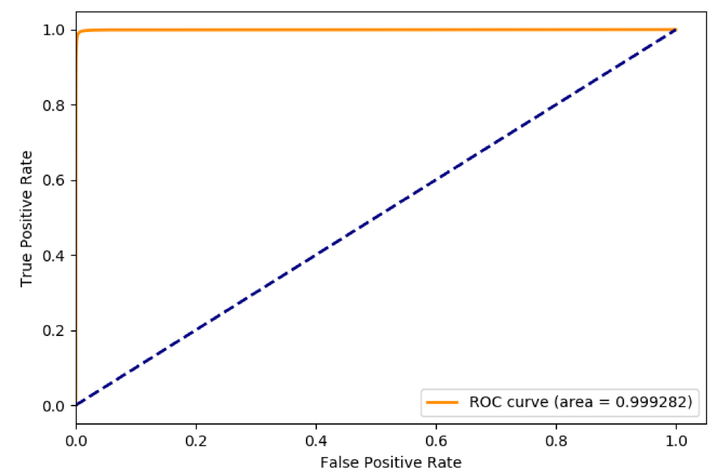


Figure 8: ROC curve for Random Forest classifier.

Conclusions

- **Random Forest surpasses** traditional PID
- **High purity sample** can be selected with **very high efficiency**
- Automation of particle selection process – **no "trial and error" attempts**
- Significantly lowered systematic uncertainties due to PID
- Quality of final classification more vulnerable to discrepancy between Monte Carlo generated and real data

Narodowe Centrum Nauki

This work has been supported by the Polish National Science Centre under agreements no. UMO-2016/21/D/ST6/01946, UMO-2016/22/M/ST2/00176, UMO-2015/19/D/ST2/01600.

Warsaw University of Technology