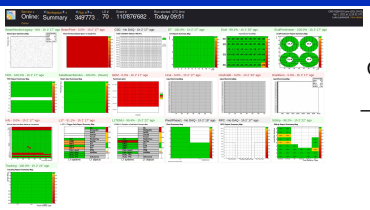
Machine learning techniques for data quality monitoring at the CMS detector

Luka Lambrecht and Rosamaria Venditti, on behalf of the CMS Collaboration
e-mail: luka.lambrecht@cern.ch, rosamaria.venditti@cern.ch, cms-PPD-conveners-DQM-DC@cern.ch

Data quality monitoring and data certification at CMS

- Goals: optimal usage of the LHC delivered luminosity.
  filter compromised data from certified data.
- Data quality monitoring: spot detector issues in real time.
- Data certification: certify data as good quality for physics analyses.
- Current manual procedure has some disadvantages:
  - Very labour intensive.
  - Sensitive to visualization details and human errors.
  - Coarse time granularity (per run instead of per luminosity section).



Current

Future

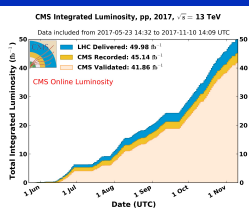


Figure from [1]

Strategies for anomaly detection

Challenges:
- No reliably labeled data for training and testing
- Large class imbalance (most data is good).
- Non-exhaustive definition of "bad" monitoring element; impossible to foresee or simulate all potential failure scenarios.

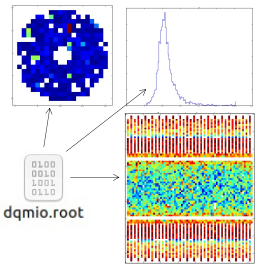
Solution:
- Use unsupervised learning.
- Employ anomaly detection methods.

Common infrastructure

Input data

Requirements:
- Similar to what is used in manual DQM/DC.
- Per luminosity section time granularity.
- Sufficiently small size on disk.
- Centrally available to the whole collaboration.

Solution: nanoDQMIO
- Subset of all DQM/DC monitoring elements.
- Per-lumisection saving.
- Available via central data aggregation system.



Preprocessing and resampling

Model training on GPU

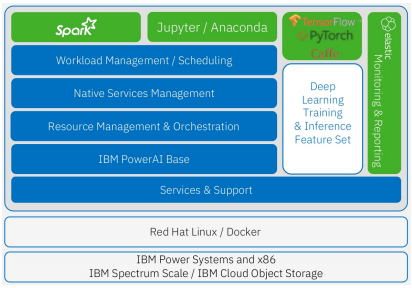


Figure from [4]

- Collaboration between CERN OpenLab and IBM.
- Large model training on GPU-equipped machines.
- Allows experimenting with large convolutional autoencoder models.

Output score summary

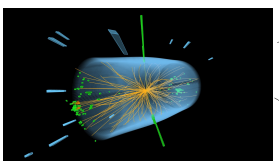
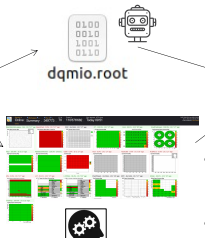
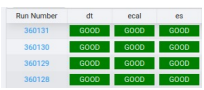


Figure from CMS-PHO-EVENTS-2020-025 (link)



[Work in progress!]



- Similar output format as manual DQM/DC in Run Registry.
- Will allow comparison and further investigation.

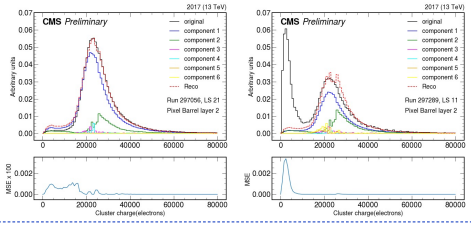
Future developments

- Further validation and commissioning in Run-3 data.
- Implement in online DQM software for live data taking (see talk from A.Harilal).
- Further tune the nanoDQMIO content.

Case study: pixel tracker

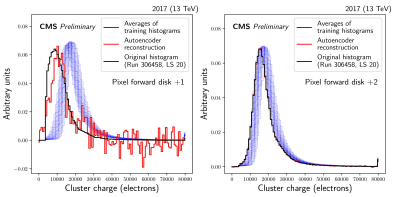
Non-negative matrix factorization

Figure from [2]



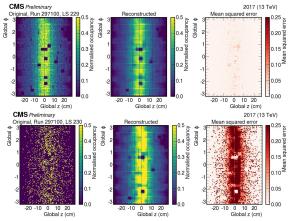
Autoencoder

Figure from [2]



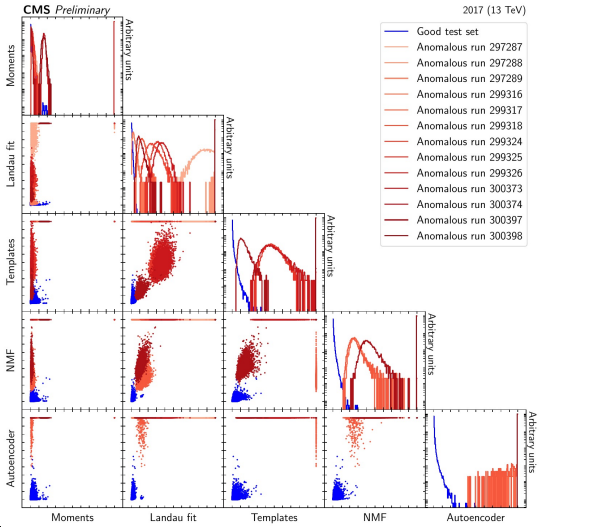
Residual neural networks

Figure from [2]



Comparison of methods

Figure from [3]



Conclusions

- Accurate automatic flagging of anomalous lumisections with sufficiently low false alarm rate.
- A few anomalies found in previously manually certified data, traced down to high voltage tests and beam dump effects.

References

[1]: CMS, *Public CMS Data Quality Information*, link
[2]: CMS, *Tracker DQM Machine Learning studies for data certification*, CMS-DP-2021-034, link
[3]: CMS, *Prospects for computer-assisted data quality monitoring at the CMS pixel detector*, CMS-DP-2022-013, link
[4]: E. Fiala, Minsky Cluster Training, IBM Corporation, 03 September 2019.

Poster presented at ACAT 2022,
23–28 October 2022, Bari, Italy
https://indico.cern.ch/event/1106990/