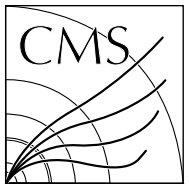
Available on CMS information server

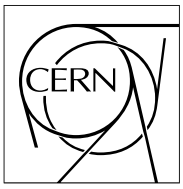
CMS DP -2024/095



The Compact Muon Solenoid Experiment

CMS Performance Note

Mailing address: CMS CERN, CH-1211 GENEVA 23, Switzerland



19 October 2024

Machine learning tools for the automatized monitoring of the CSC detector

CMS Collaboration

Abstract

Ensuring the quality of data in large HEP experiments such as CMS at the LHC is crucial for producing reliable physics outcomes, especially in view of the high luminosity phase of the LHC, where the new data taking conditions will demand a much more careful monitoring of the experimental apparatus. The CMS protocols for Data Quality Monitoring (DQM) are based on the analysis of a standardized set of histograms offering a condensed snapshot of the detector's condition. In the standard approach the CMS DQM has a per-run time granularity. Recently, unsupervised machine learning models have been successfully employed for the anomaly detection in DQM with per-lumisection granularity, with the goal to improve the certification of the data collected in Run3 and the online monitoring of the detector. In this note, we discuss the development of an automated workflow for the online DQM of the Cathode Strip Chambers, based on a ResNet autoencoder trained on CSC occupancy maps. The data pre-processing will be illustrated, together with a description of the training procedure and a first validation on 2024 data.

Machine learning tools for the automatized monitoring of the CSC detector

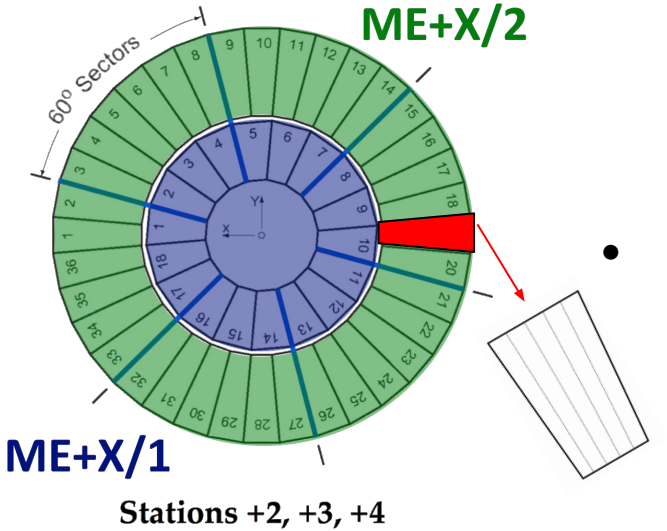
CMS Collaboration
cms-dpg-conveners-csc@cern.ch

Definitions and terminology

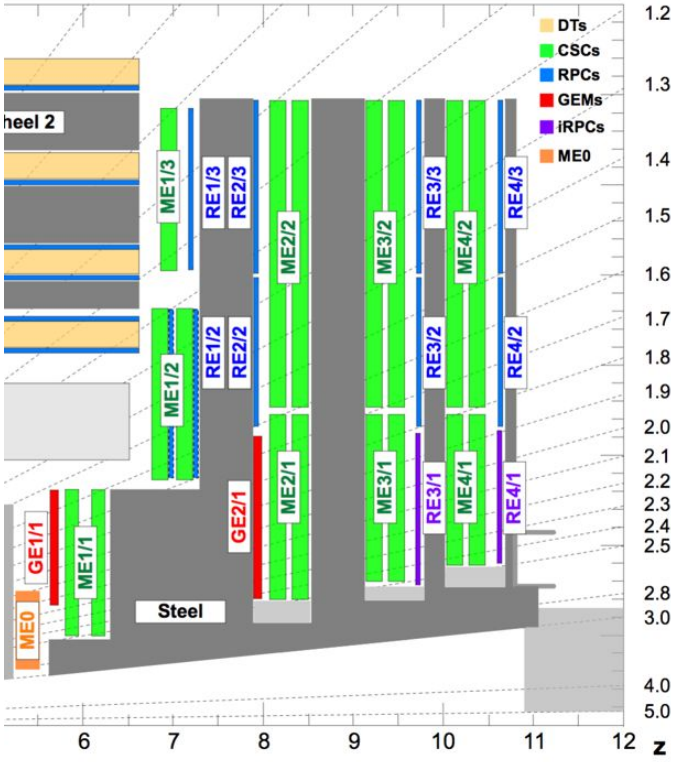
- **Fill:** A period during which the same proton beams are circulating in the LHC, typically spanning multiple runs. Every time the LHC injects beams in the machine, it marks the beginning of what is known as a Fill.
- **Run:** A time unit of data taking in CMS, typically consisting of a few tens to a few hundreds of luminosity sections.
- **Luminosity section (LS):** An elementary time unit of continuous data taking in CMS, during which the instantaneous luminosity is assumed unchanged. A LS lasts 218 LHC orbits, or approximately 23.3 seconds.
- **Data quality monitoring (DQM):** The process of checking the quality of recorded data, with the aim of spotting potential detector issues.
- **Data certification (DC):** The process of checking the quality of recorded data, aiming to certify the data as good for usage in physics analyses
- **F1 score:** in machine learning, the F1 score is an evaluation metric commonly used in classification tasks, combining the precision (p = TP/(TP+FP)) and recall (r = TP/(TP+FN)) as follows: F1 = 2p*r/(p+r), where TP = True Positives, FN = False Negatives, FP = False Positives.
- **StreamExpress**: This datastream is collected with a suite of different triggers, some of them involving muons. The set of triggers used is similar to the one used in the online DQM.

Definitions and terminology

- Cathode Strip Chambers (CSCs) are placed in the endcaps of the CMS muon system [1]
- Chambers are arranged in disks called stations, four on each side, numbered from ME-4 to ME+4



- The stations ME±1 have three rings (ME±1/1, ME±1/2, ME±1/3), while the other stations have two rings (ME±X/1 and ME±X/2, where X=2...4)
- ME±X/1 chambers span 20°, ME±X/2 chambers span 10° and both types are readout by 5 frontend boards.



Context and Motivations

- Machine Learning (ML) tools are widely used for Anomaly Detection tasks at CMS, i.e. identifying malfunctions in the detector as well as monitoring the quality of the data

- Other CMS subsystems already developed ML-based tools both for DC [2] and Offline DQM [3], as well as for the online DQM [4] and a joint effort is ongoing to coordinate the different developments and provide common platforms and tools under the PPD/DQMDC/ML4DQM group.

- In particular, the tool developed by ECAL [4], presently integrated in the CMS software (CMSSW), shows promising and stable performance for monitoring the detector with a per-LS time granularity.

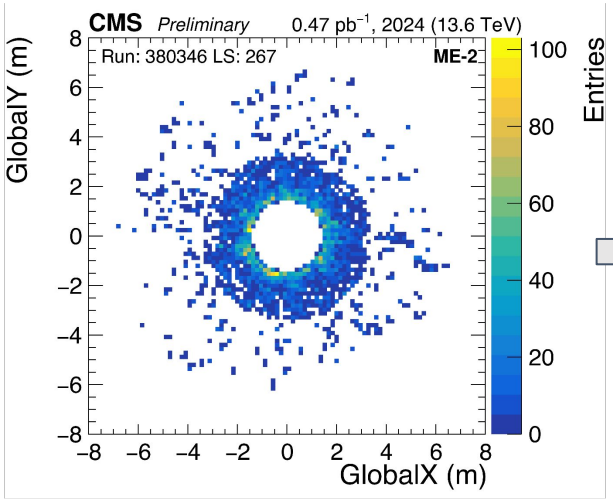
- In this context, we explore for the first time the applicability of the method developed by ECAL to the Online monitoring of the CMS Muon System. The development presented in this note uses CSC occupancy plots to showcase the feasibility of a fine time granularity monitoring of the Muon System.

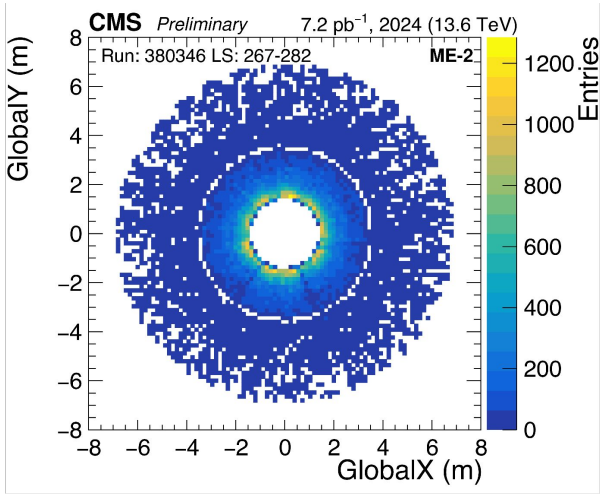
Method

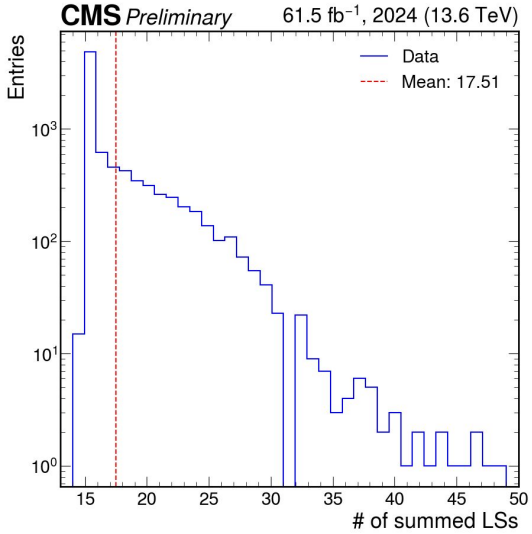
- The **dataset** used is made of 150k LSs certified as GOOD in the StreamExpress dataset collected in 2024. We use CSC occupancy plots, each showing the distribution of reconstructed hits over the detector surface as implemented in the Offline DQM, but with a per-LS time granularity.
- The **ML architecture**, mostly taken from [4], is based on a ResNet Auto-Encoder (AE) trained on the aforementioned CSC occupancy plots.
- The reconstruction quality is quantified by the loss, i.e. the difference between an occupancy plot and its autoencoder reconstruction, computed using the L1Loss. Anomalous images will differ from their reconstruction, thus leading to high values of the loss.
- While this method applies for the monitoring of any subdetectors, the preprocessing of the data and the figure of merit (FOM) for performance evaluation benefit from a prior knowledge of the **most common expected anomalies**, which are detector specific. For the CSC use-case, anomalies are expected to look like lack or excess of entries in a chamber (20° or 10°) or in a readout sector (2° sector along the ϕ coordinate).
- The occupancy values depend on the **instantaneous luminosity and on the η coordinate**. Therefore, we **aggregate** the input images based on the luminosity (integrated luminosity > 7pb-1) and we **rebin** the low-η regions to compensate for the low occupancies, while the loss map is **corrected** to be uniform as a function of η. Moreover, two models for the inner and outer CSC wheels are trained and optimized independently.
- In order to be sensitive to both types of anomalies, we use the loss value with its sign. For each image we use the maximum and minimum value of the loss as FOM. We therefore optimize two **thresholds** to maximize the flagging of known anomalies while minimizing the false alarm rate.

Data pre-processing: merging consecutive LS before training

Left: 2D map of the reconstructed hit position (x vs y) in CSC ME-2 for one single lumisection (LS 267 Run 380346, collected on May 03, 2024), showing low occupancy in the low η region.
Center: 2D map of the reconstructed hit position (x vs y) in CSC ME-2 obtained summing LS from 267 to 282 Run 380346, corresponding to an integrated luminosity of 7.2 pb-1.
Right: Distribution of the number of LS that have been combined to form the images used for training over the entire input dataset, corresponding to an integrated luminosity of 61.5 fb-1.



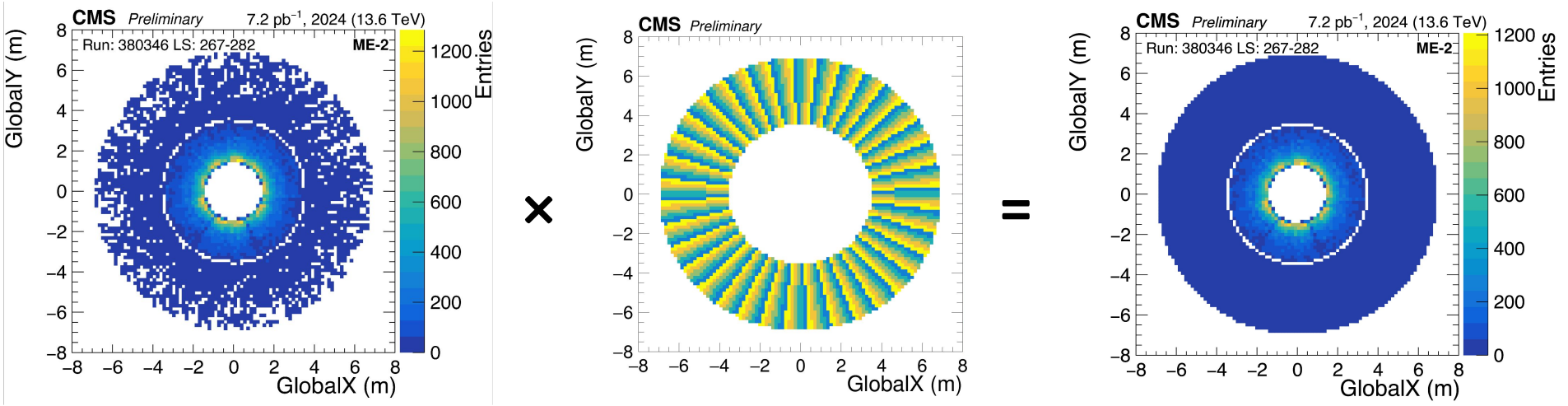




Data pre-processing: rebinning

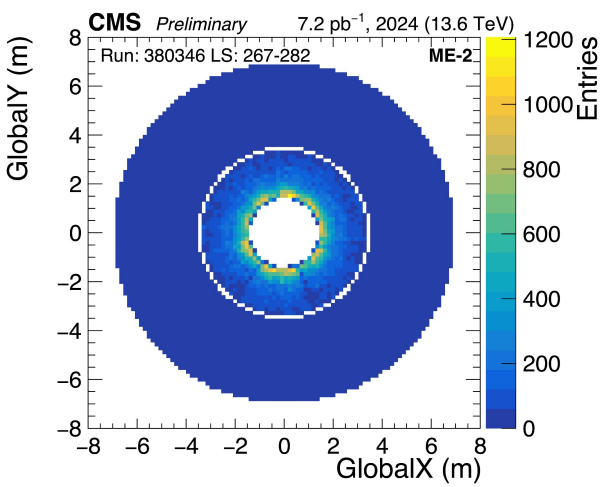
After merging consecutive LS, the images are still characterized by empty regions at low η. The statistical fluctuations can affect the generalization capability of the autoencoder. Therefore, following the geometry of the detector readout, the external part of each image is rebinned in slices along the ϕ coordinate. For each slice, the color corresponds to the average number of entries computed in the original image.

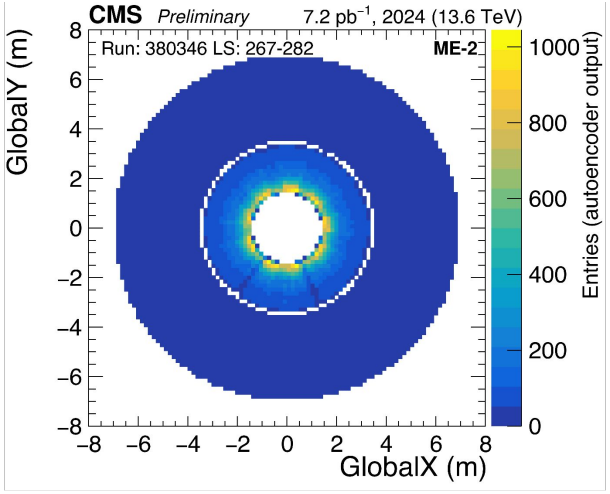
Center: mask showing the geometry used for rebinning. **Right:** 2D map of the reconstructed hit position (x vs y) in CSC ME-2 after the rebinning corresponding to 15 consecutive LS (from 267 to 182) in Run 380346, collected on May 03, 2024.

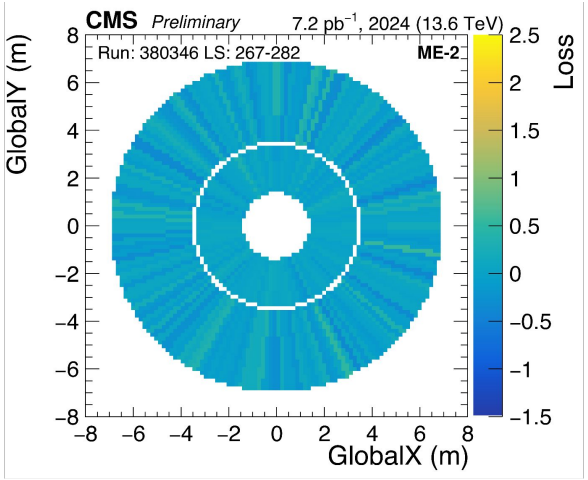


Autoencoder training

The autoencoder is **trained on a set of unlabeled images from certified data** collected in 2024, assuming that the large majority of the examples used for the training represents a properly functioning detector. The training is unsupervised. After training, the algorithm is able to reconstruct "good" occupancy plots. Here we show the results of the training for the ME-2/1 and ME-2/2 stations. While two algorithms are **trained separately on ME-2/1 and ME-2/2**, here we show the combined images. **Left:** image used in the training, extracted from good certified data after preprocessing, where lower occupancies in two frontend boards in the ME-2/1 station can be seen in all plots and were therefore learned by the algorithm. **Center:** the image as reconstructed by the algorithm. **Right:** 2D map of loss computed as the signed difference between input and reconstructed image.





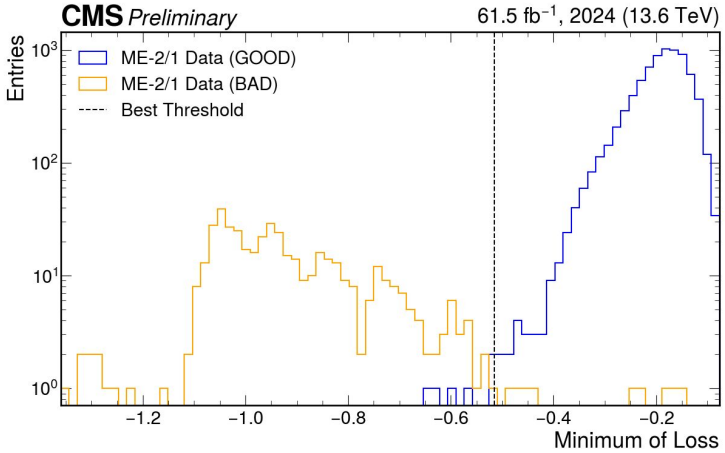


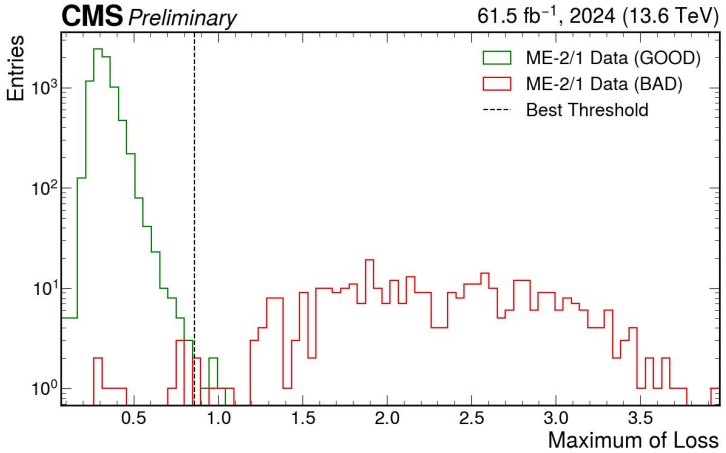
Definition of anomalies (ME-2/1)

While data certified as good is expected to mostly contain “normal” images, transient issues which don’t prevent the data to be certified as “good” can be present. For the ME-2/1 station, a set of good and problematic images is isolated from data by visual inspection. Examples are shown in the next pages. For the two datasets (“bad” and “good”), we look at the distribution of the minimum (**left plot**) and maximum (**right plot**) values of the loss, where one entry of each histogram corresponds to one image. We then set two decision thresholds, maximizing the F1 score for each distribution.

Confusion Matrix

		Actual	
		True	False
Predicted	True	99.0 %	0.1 %
	False	1.0 %	99.9 %



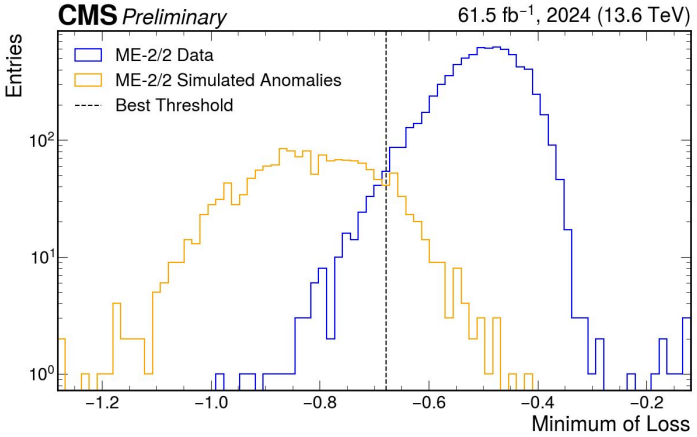


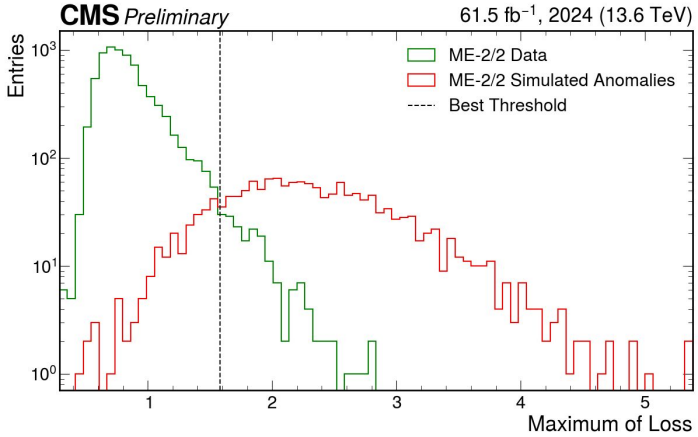
Definition of anomalies (ME-2/2)

Contrary to what done for ME-2/1, in this case anomalous images are difficult to isolate by eye because of the low occupancy characterizing the outer region and the statistical fluctuations. Therefore, we generate a set of anomalous images by artificially introducing over fluctuations and under fluctuations of the entries in some slices of the detector, based on the geometry of the readout sectors and based on few anomalous instances found in real data. Examples are shown in the next pages. For the two datasets ("bad" and "good"), we look at the distribution of the minimum (**left plot**) and maximum (**right plot**) values of the loss, where one entry of each histogram corresponds to one image. We then set two decision thresholds, maximizing the F1 score for each distribution.

Confusion Matrix

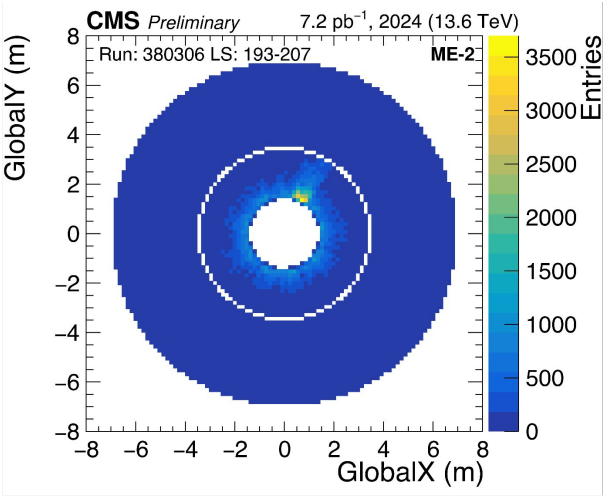
		Actual	
		True	False
Predicted	True	87.6 %	4.8 %
	False	12.4 %	95.2 %

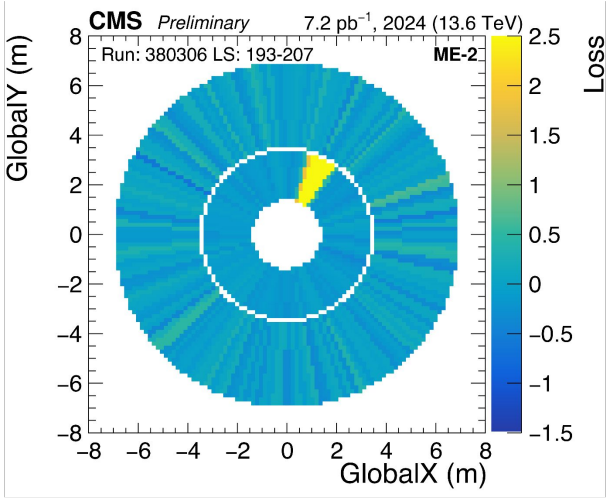


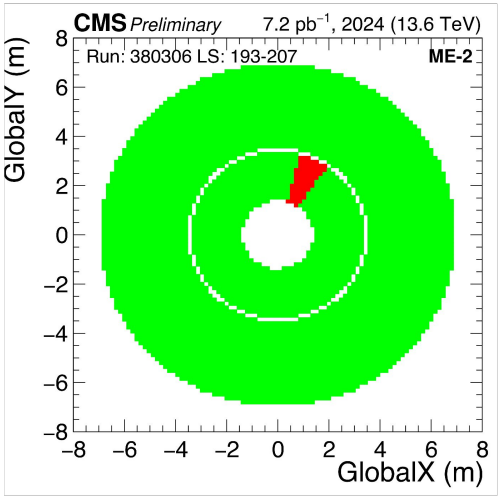


Results: example of a anomalous image in ME-2/1 from data and its corresponding loss map

For the ME-2/1 station, a set of problematic images is isolated from data by manual inspection in order to evaluate the confusion matrix. **Left:** Input image for LS range 193 to 207 in Run 380306 collected on May 02, 2024, showing high occupancy in one chamber. **Center:** The loss map is computed as the signed difference between the input and the reconstructed images, which is then normalised to be uniform along the η coordinate. **Right:** a map is provided clearly showing the anomalous region, defined based on the two decision thresholds on the maximum and minimum values of the loss across the 2D map.

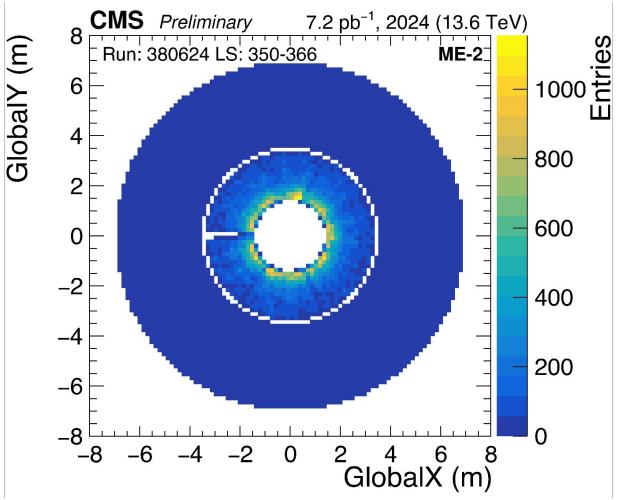


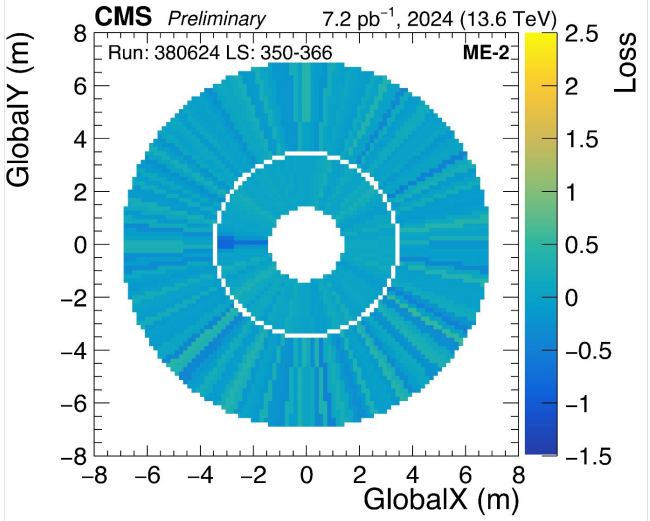


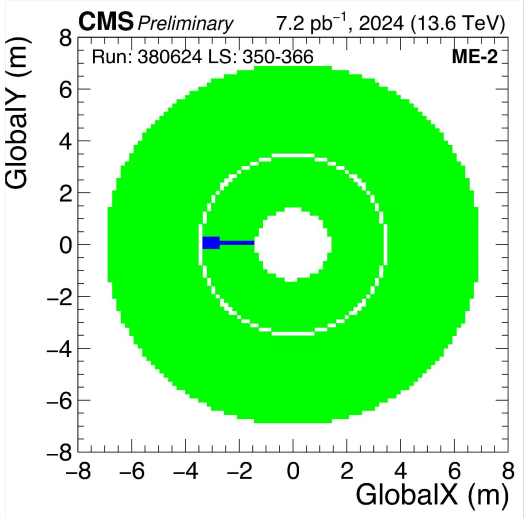


Results: example of a anomalous image in ME-2/1 from data and its corresponding loss map

For the ME-2/1 station, a set of problematic images is isolated from data by manual inspection in order to evaluate the confusion matrix. **Left:** input image for LS range 350 to 366 in Run 380624 collected on May 11, 2024, showing low occupancy in one readout sector. **Center:** The loss map is computed as the signed difference between the input and the reconstructed images, which is then normalised to be uniform along the η coordinate. **Right:** a map is provided clearly showing the anomalous region, defined based on the two decision thresholds on the maximum and minimum values of the loss across the 2D map.

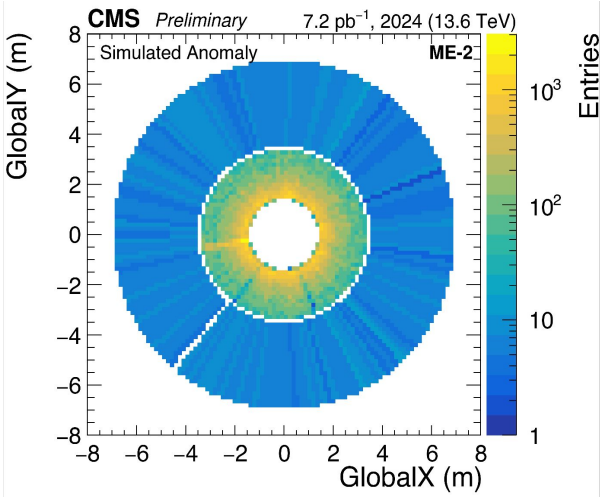


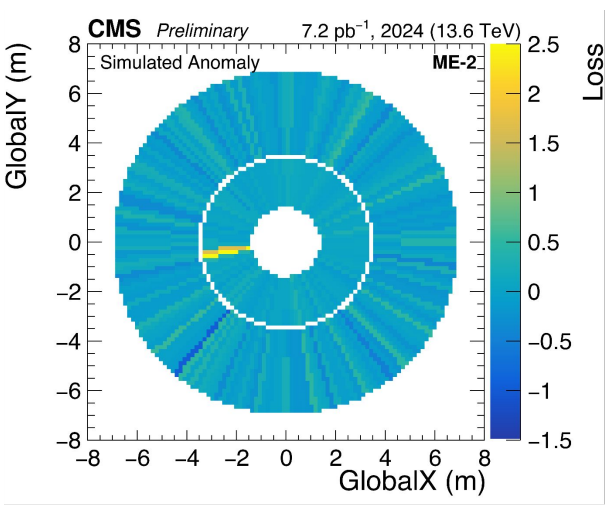


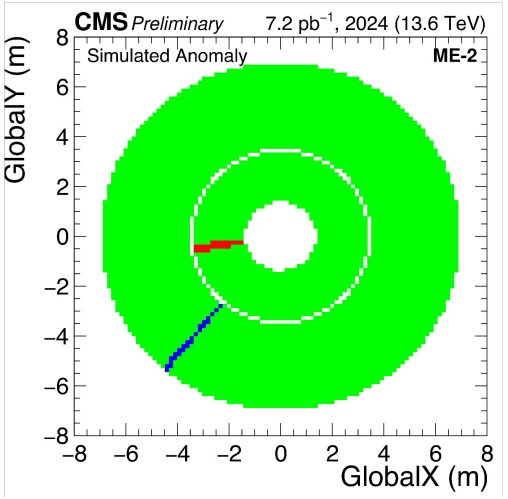


Results: example of a simulated anomalous image in ME-2 and its corresponding loss map

For the ME-2/2 station, a set of problematic images is generated starting from real data and introducing some over fluctuations and under fluctuations of the entries in some slices of the detector. These generated images are used to evaluate the confusion matrix. **Left:** simulated input image showing low occupancy in one readout sectors of ME-2/2 and high occupancy in one readout sector of ME-2/1. The image is created starting from an averaged image from data. Given the low occupancy values at low η, the z axis is displayed in log-scale. **Center:** The loss map is computed as the signed difference between the input and the reconstructed images, which is then normalised to be uniform along the η coordinate. **Right:** a map is provided clearly showing the anomalous regions, defined based on the two decision thresholds on the maximum and minimum values of the loss across the 2D map.



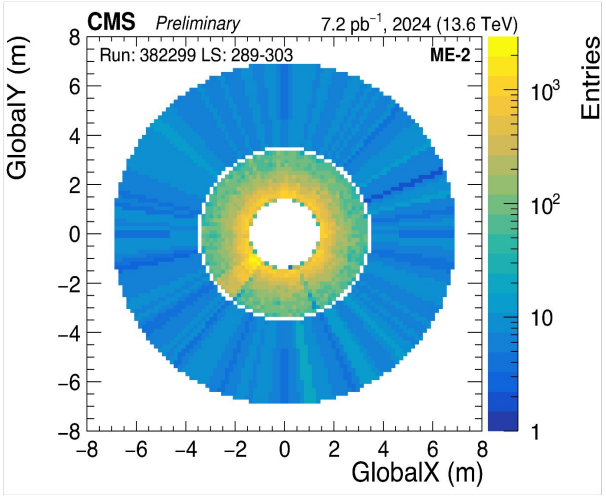


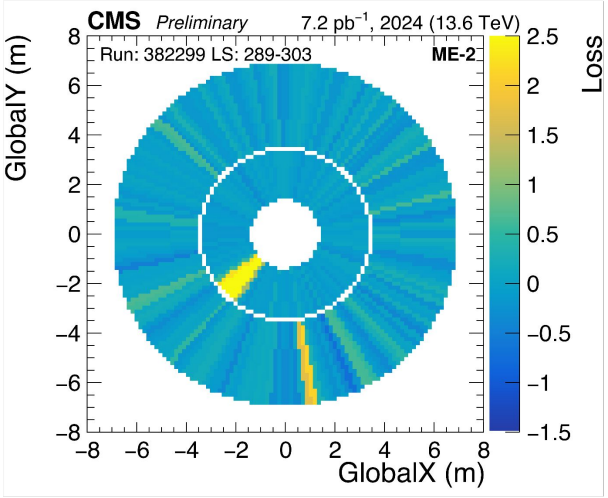


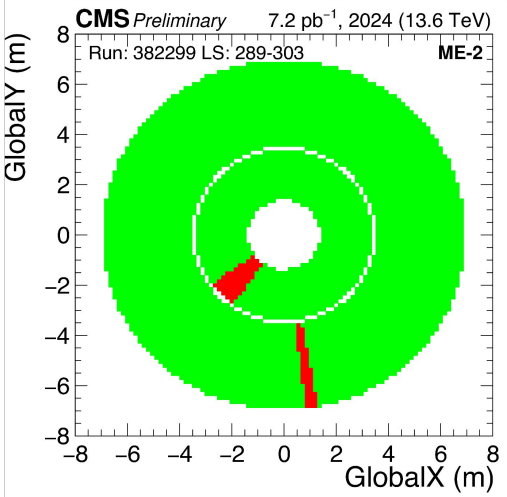
Results: example of a anomalous image in ME-2 from data and its corresponding loss map

In real data, anomalous images which were neither labeled by manual inspection nor generated can be identified based on the maximum (minimum) value of the loss being above (below) threshold.

Left: input image for LS range 289 to 303 in Run 382299 collected on June 21, 2024, showing high occupancy in one ME-2/1 chamber and in one ME-2/2 readout sector. **Center:** The loss map is computed as the signed difference between the input and the reconstructed images, which is then normalised to be uniform along the η coordinate. **Right:** a map is provided clearly showing the anomalous regions, defined based on the two decision thresholds on the maximum and minimum values of the loss across the 2D map.







Summary

- An anomaly detection tool has been developed for the online monitoring of the CSC detector with time granularity of the order of few lumisections, to showcase the applicability of ML-based monitoring to the CMS Muon System.

- Anomalies are identified in the 2D occupancy maps of reconstructed hits in the CSC ME2-4 stations. In this note, results for the ME-2 station are reported.

- The spatial granularity of the tool is determined by the available 2D map already implemented in the DQM. In order to control the statistical fluctuations in the outer disks, a rebinning is applied which follows the geometry of the readout sectors.

- The algorithm is trained on a set of images from certified data, each corresponding to about 15 LS, separately for ME-2/1 and ME-2/2 stations. After training, a strategy is defined for identifying anomalous images. The fraction of anomalous images correctly labeled by the algorithm is above 85% for both the ME-2/1 and ME-2/2.

Bibliography

[1] Layter, J. G., "*The CMS muon project: Technical Design Report*". CERN (1997) https://cds.cern.ch/record/343814

[2] CMS Collaboratiom, *"CMS tracker data quality certification with new machine learning tools",* CERN-CMS-DP-2024-070 (2024), https://cds.cern.ch/record/2905834

[3] *CMS Collaboration, "Machine Learning Techniques for JetMET Data Certification of the CMS Detector", CERN-CMS-DP-2023-032 (2023) https://cds.cern.ch/record/2860924*

[4] The CMS ECAL Collaboration, *"Autoencoder-Based Anomaly Detection System for Online Data Quality Monitoring of the CMS Electromagnetic Calorimeter"*, *Comput Softw Big Sci* **8**, 11 (2024), https://doi.org/10.1007/s41781-024-00118-z