PROCEEDINGS OF SCIENCE

AutoEncoders for per-lumisection data quality monitoring at CMS

Alkis Papanastassiou[a,*] **and Valentina Gori**[b] **on behalf of the CMS collaboration**

[a] *INFN, Sezione di Firenze - Firenze, Italy*

[b] *Baker Hughes (Nuovo Pignone Tecnologie), Via Felice Matteucci 2, 50127 Firenze, Italy*

E-mail: alkis.papanastassiou@cern.ch, valentina.gori@bakerhughes.com

The monitoring of data quality is crucial both online, during the data taking, to promptly spot issues and act on them, and offline, to provide analysts with datasets that are cleaned against the occasional failures that may have crept in. Typically, data quality monitoring (DQM) is performed by *shifters* who look at a set of integrated quantities, compare them with reference histograms, and, based on their experience and training, assign quality flags. Recently CMS has developed the possibility of producing DQM plots per-lumisection, where a lumisection is a time unit corresponding to about $23s$ of data taking. To analyze per-lumisection data, a manual approach would be prohibitive due to the high number of lumisections, therefore an automated one would be preferable. In this work, the first use in CMS of AutoEncoders to perform anomaly detection on per-lumisection data, specifically for quantities associated with jets and missing transverse energy, is presented. The technique developed allows the detection of anomalies at the level of individual lumisections, which might be overlooked when examining integrated quantities, and serves as a proof of concept regarding the efficacy of this and similar approaches.

The Eleventh Annual Conference on Large Hadron Collider Physics (LHCP2023)
22-26 May 2023
Belgrade, Serbia

[*] Speaker

PoS(LHCP2023)284

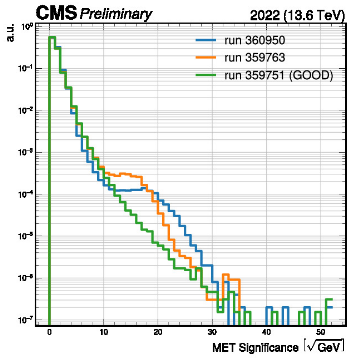


Figure 1: Histograms of a Monitor Element (MET Significance) for three different runs, one flagged *GOOD* and two presenting an anomaly, therefore flagged *BAD*.

1. Introduction

In CMS [1], Data Certification (DC) is the final step of quality checks performed by Data Quality Monitoring (DQM) on recorded collision events. Data are gathered in luminosity sections, *lumisections* in short (LSs), corresponding to $\sim$ 23 seconds of data taking. LSs are grouped in runs. For each run, experts monitor several reconstructed distributions called Monitor Elements (MEs) to spot issues and, or, anomalies, in the data. For the specific case of quantities pertaining to hadronic jets and missing transverse momentum (MET), an issue in a few LSs would cause the entire run to be flagged as problematic (*BAD*), and thus removed from the pool of *good-for-analysis* data (*GOOD*). In Fig.1, the histogram illustrating a specific ME (MET Significance) is presented for three distinct runs— one categorised as *GOOD* and the other two as *BAD*.
MET Significance is defined as:

$$\text{METSig} \equiv \frac{\text{MET}}{\sqrt{\text{SumET}}} = \frac{MET}{\sqrt{\sum |\vec{p}_T|}} ,$$

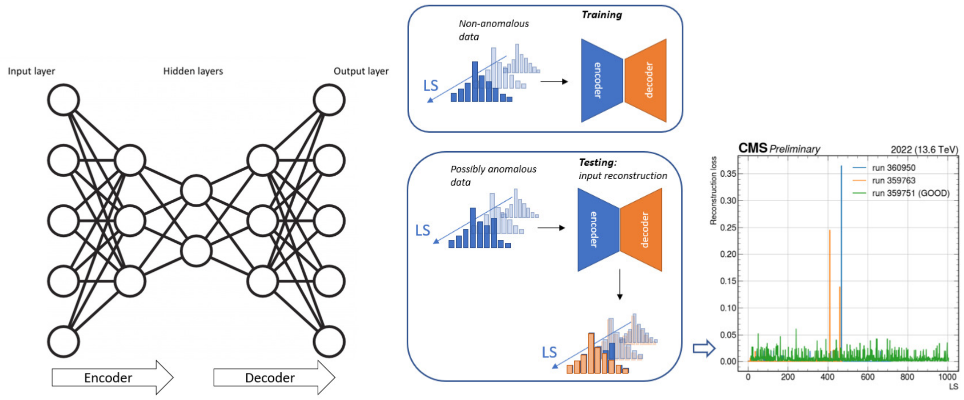
with $\vec{p}_T$ denoting the transverse momenta of all the reconstructed objects.

2. Per-LS data

In CMS, the possibility of accumulating quantities monitored for data quality purposes per-LS has been recently extended to Jet and Missing Energy (JME) MEs. This possibility allows for a higher granularity detection of anomalies, potentially enabling the saving of higher amounts of data from runs presenting only a limited set of anomalous LSs. Given the high number, $O(1000)$, of LSs to be analysed for each run, an automated approach (rather than a manual one) for DC is required. Machine Learning (ML), particularly Neural Networks (NN) [2], can be implemented to this end. An unsupervised ML model based on a specific NN architecture called AutoEncoder (AE)[3] is employed.

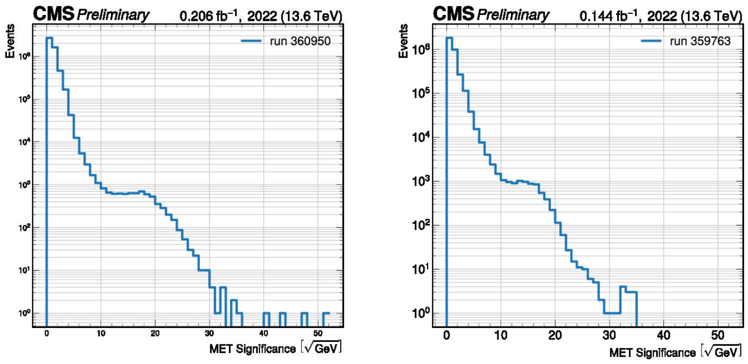
3. AutoEncoder-based Anomaly Detection Tool

The model that was optimised is an Under-complete AE [3] built using dense layers with three hidden layers in total, see Fig.2a. The AE is trained on non-anomalous data from *GOOD* runs:



(a) **(b)**

Figure 2: Structure of the Under-complete AE (the number of nodes is just indicative) (a) and scheme of training and testing steps for the model (b).



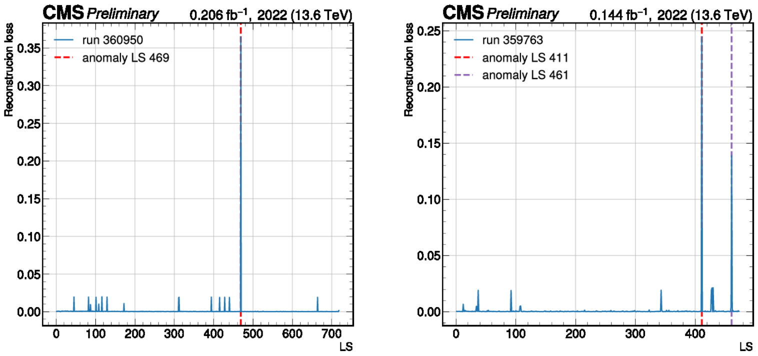
(a) Run 360950. **(b)** Run 359763.

Figure 3: Histograms of METSig for the two runs under study, an anomalous shape is evident in both.

histograms of specific MEs are fed to the model with per-LS granularity to allow the AE to learn a *normal* non anomalous behaviour of that specific ME, see Fig.2b. The training is performed via the minimisation of the reconstruction loss, a measure of the distance between the input and output of the AE. In this case the reconstruction loss is the mean squared error, MSE = $\frac{1}{\mathrm{n}}\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{y_i}-\hat{y}_i)^2$, where y and $\hat{y}$ are respectively the input and the output of the AE and n is the bin number. Possibly anomalous runs under investigation are tested by looking again at the reconstruction loss: peaks in this function indicate LSs containing histograms that deviate from the learned behaviour.

4. Results

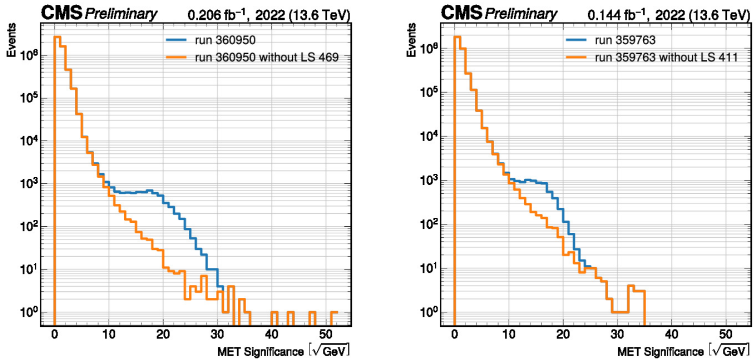
The model was tested on two runs, 360950 and 359763, that where flagged *BAD* by JME due to the presence of an anomaly visible in histograms of many different MEs, see e.g. Fig.3. By analysing the per-LS MET Significance for both runs via the AE-based anomaly detection tool, peaks were observed in the reconstruction loss limited to a small number of LSs. This can be seen in Fig.4, which shows run 360950 presenting a peak corresponding to LS 469 and run 359763 displaying two peaks, the biggest one corresponding to LS 411, and the smaller one corresponding to LS 461. Once anomalous LSs are identified they are removed from the run. The resulting



(a) Run 360950. **(b)** Run 359763.

Figure 4: Reconstruction loss for the two runs under study showing the presence of a limited set of high peaks.

histograms for both *BAD* runs show how the cause of the MET Significance anomaly was LS 469 for run 360950 and LS 411 for run 359763, see Fig.5. The exclusion of LS 461 results in a smoothing of the tail of the histogram, consequently rendering this particular LS not attributable as the origin of the anomaly.



(a) Run 360950, with and without LS 469. **(b)** Run 359763, with and without LS 411.

Figure 5: The two runs with and without the LSs responsible of the anomaly.

5. Conclusions

An AutoEncoder-based Anomaly Detection Tool capable of detecting anomalies in DQM MEs with a per-LS granularity has been developed. The tool was tested on several runs flagged *BAD* by JME DQM and identified the source of the anomalous behaviour in a limited set of LSs. In particular, one LS was removed from each run presented in this work and it was verified that the remainder was no more anomalous. The equivalent luminosity recovered from the two runs is $\sim 350\, pb^{-1}$. Exploiting the per-LS granularity in DQM and systematically employing the tool that was presented, will enable an increase in the efficiency of the DC procedure, ultimately resulting in a larger dataset available for physics analyses.

This work uses results that are part of a CMS Detector Performance Note (DP-note)[4].

References

[1] The CMS Collaboration et al. *JINST*, **3**, 2008, S08004.

[2] Goodfellow I. et al. *Deep Learning*. MIT Press, 2016.

[3] Hinton, G. E. and Salakhutdinov, R.R., *Reducing the Dimensionality of Data with Neural Networks, Science*, **313**, 5786 , 504-507.

[4] CMS Collaboration, *An AutoEncoder-based Anomaly Detection tool with a per-LS granularity,* CMS DP-2023/010.