

Machine Learning Techniques for JetMET Data Certification of the CMS Detector

J. ALTORK

ON BEHALF OF THE CMS COLLABORATION

Department of Physics and Astronomy, Via G. Sansone 1, 50019 Sesto Fiorentino (FI), Italy

INFN Firenze, Via G. Sansone 1, 50019 Sesto Fiorentino (FI), Italy

Summary. — In CMS, data quality monitoring and data certification are crucial for ensuring reliable data suitable for physics analyses. In the offline DQM procedure, the quality of recorded data, grouped in runs, is evaluated. The current method for certifying quantities related to hadronic jets and missing transverse momentum relies mainly on manually monitoring reference histograms that summarize detector status and performance. Due to the large number of distributions, this process is time intensive and prone to human error, especially when deviations are subtle. This work presents machine learning methods for certifying offline DQM data, focusing on hadronic jet and MET objects. Using 2018 and 2022 collision data, we show that autoencoder techniques can accurately certify runs and detect ineffective detector regions, reducing both certification time and the chance of missing anomalies.

1. – Introduction

In the CMS experiment [1], Data Quality Monitoring (DQM) and Data Certification (DC) [2] are essential processes that ensure the collected data is reliable and suitable for physics analyses. DQM is used to monitor the performance of the detector and the quality of recorded data, while DC is the step where experts certify which data can be used for analysis. These tasks are performed at two levels: online, which evaluates data during data-taking, and offline, which reviews data after collection, grouped into larger time intervals called Runs.

During the offline DQM procedure, the quality of the data is assessed by experts using summary plots and reference histograms that reflect the status and performance of the detector. For quantities related to hadronic jets and missing transverse energy (MET), the current certification approach relies primarily on manually checking a large number of distributions. This method is time-intensive and susceptible to human error, especially when deviations from normal behavior are subtle and difficult to spot.

To improve this process, machine learning (ML) techniques are used to automate parts of offline DQM and reduce dependence on manual certification. In particular, unsupervised ML models like autoencoders are capable of learning normal detector behavior and identifying anomalous patterns in the data.

This study presents the application of autoencoder-based methods for certifying offline DQM data, focusing on hadronic jet and MET observables [3]. Using collision data from 2018 and 2022, we show that these models can accurately certify Runs and identify problematic detector regions.

2. – Variable reduction

Before training the data with autoencoders, the number of input variables are reduced to simplify the model and focus on the most effective observables in detecting anomalies. To do so, supervised learning methods were applied to a dataset of already certified Runs.

A fully connected neural network classifier was trained with the following setup:

- Input: Mean values of 124 jet-related monitoring distributions for each Run from the 2018 dataset.
- Output (labels): Binary classification of each Run as GOOD or BAD, based on expert certification.

The dataset was split into training (70%) and testing (30%) subsets, each containing a balanced mix of good and bad Runs. The model was trained to minimize the binary cross-entropy loss, defined as

$$(1) \quad \mathcal{L} = -\frac{1}{n} \sum_{i=1}^n [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)],$$

where n is the number of samples, y_i is the true label, and \hat{y}_i is the predicted probability for sample i .

To determine which input features contribute most to the prediction, the first-order gradient of the loss with respect to each input variable was computed. This measures the model's sensitivity to changes in each feature. As shown in Figure 1, the features with the largest absolute gradients, mostly jet energy fractions, contribute most to the prediction. This confirms that jet energy fractions are the most effective inputs for data certification, and thus are used exclusively in the autoencoder models described in the following sections.

3. – Data Certification with Supervised Classification

To evaluate the effectiveness of the selected jet energy fraction observables in detecting potentially anomalous Runs, we tested three different classifiers: K-Nearest Neighbors (KNN), Gaussian Naive Bayes (Gaussian-NB), and Support Vector Machine (SVM).

The dataset used for this study consists of the mean values of 25 jet energy fraction variables. These include observables such as Charged Hadron Energy Fraction, Neutral Hadron Energy Fraction, Photon Energy Fraction, Hadron Energy Fraction, and Electron Energy Fraction in the Forward region, as well as the Neutral Constituents Fraction. Each observable is considered separately for the Barrel and EndCap detector regions and across low, medium, and high transverse momentum ranges.

As shown in Figure 2, the predicted probability histograms for good (blue) and bad (red) Runs are presented for both training and test sets across the three classifiers.

80 mixture of good and bad Runs. The model is trained using the mean squared error
 81 (MSE) as the loss function, defined as:

$$(2) \quad \text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

82 where y_i is the true input value and \hat{y}_i is the reconstructed output.

83 The anomaly detection strategy is as follows:

- 84 1. Compute the maximum MSE value across all training good Runs and set it as a
 85 threshold.
- 86 2. For each test Run, compute the maximum MSE across all features.
- 87 3. If the maximum MSE of a test Run exceeds the threshold, the Run is classified as
 88 bad; otherwise, it is classified as good.

89 Two datasets were used for evaluation. In the first case, both training and testing
 90 were performed using Runs from 2018 collision data. In the second case, the model was
 91 trained on a combination of Runs from 2018 and 2022 and tested on separate Runs from
 92 2022. The autoencoder showed strong performance in both setups, achieving over 90%
 93 accuracy in certifying the Runs. However, only the results from the second dataset are
 94 presented here.

95 Figure 3 shows the mean squared error of good training Runs from 2018 and 2022
 96 across the 25 jet energy fraction features. The red dashed line indicates the threshold
 97 applied to test Runs. Then, Figures 4a and 4b show the maximum mean squared error of
 98 each Run in training and testing, using a threshold of 0.002 derived from the maximum
 99 MSE of the good training Runs. As shown in the plots, the autoencoder clearly separates
 100 good and bad Runs: good test Runs fall below the threshold, while bad ones exceed it,
 101 demonstrating the model's effectiveness in detecting anomalous detector behavior.

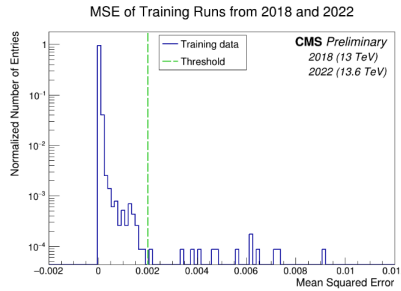


Fig. 3. – Mean squared error of training good Runs from 2018 and 2022 across the 25 jet energy fraction features. The red dashed line indicates the selected anomaly detection threshold.

102 5. – Anomaly Detection with Autoencoder

103 The structure of the autoencoder was modified in this part of the study to operate di-
 104 rectly on entire histograms, rather than on mean values. This allows the model to capture
 105 more detailed patterns and localized anomalies within the distributions themselves.

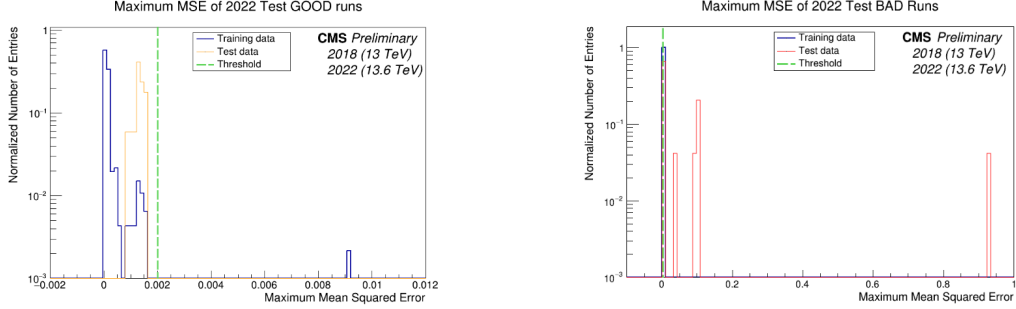


Fig. 4. – (a) Maximum mean squared error per Run, training good Runs (blue) and test good Runs (orange). (b) Maximum mean squared error per Run, training good Runs (blue) and test bad Runs (red). The horizontal green line shows the threshold of 0.002.

5.1. MET histograms. – In this study, MET histograms are used to investigate specific Runs that exhibit MET tails, which are known to be problematic for certain analyses. A fully connected autoencoder is trained on a dataset of 2018 good Runs and is then applied to bad Runs suffering from MET tails in order to detect deviations from expected behavior.

For this AE, the mean absolute error (MAE) is used as the loss function:

$$(3) \quad \text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

where y_i is the true bin value in the histogram, and \hat{y}_i is the reconstructed bin value.

The maximum MAE observed across the training good Runs is used as a threshold. Any test Run whose reconstruction loss exceeds this threshold is flagged as potentially anomalous due to MET tails.

Figure 5 and Figure 6 illustrate the performance of the autoencoder in detecting MET tails. Figure 5 shows the MET histograms for a good and a bad Run, comparing the original input with the reconstructed output. Figure 6 shows the MAE loss between the input and reconstructed MET histograms, shown on a logarithmic scale. As seen in the plots, the model detects significantly higher reconstruction loss for the bad Run, confirming the presence of MET tails.

5.2. 2-D jet histograms. – The goal of this study is to evaluate and differentiate detector efficiencies across different data taking period and detector configuration.

The autoencoder is trained on several types of 2-dimensional histograms:

- Phi vs Eta
- Hadron Occupancy
- Neutral Hadron EndCap Occupancy

The mean squared error (MSE) is used as the loss function in this model:

$$(4) \quad \text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

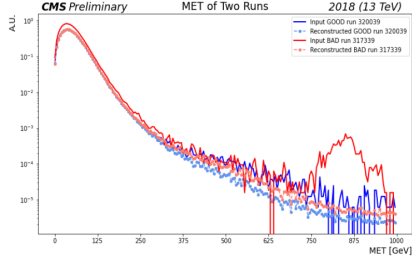


Fig. 5. – MET histograms for a good Run 320039 and a bad Run 317339, showing both the original input and the reconstructed output distributions.

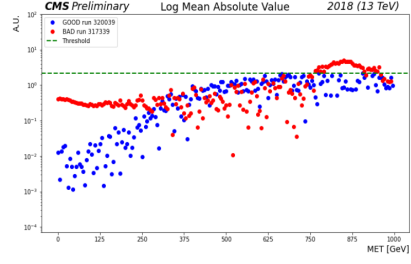


Fig. 6. – MAE loss between input and reconstructed MET histograms, plotted on a logarithmic scale.

129 The model is trained on Runs from 2018 and tested on different Runs from the same
 130 year to evaluate its sensitivity to detector configuration changes. It performs well in
 131 both cases: accurately reconstructing histograms from the same period and detecting
 132 anomalies when applied to data from a different period.

133 Figures 7–9 show the result for Run 316114, using data from the same period for
 134 training and testing. The original histogram, reconstruction, and loss map indicate
 135 minimal error across the detector. Then, Figures 10–12 show the result for Run 320712,
 136 from a different period. The loss map reveals a clear anomaly in the barrel region,
 137 indicating a lost sector. This confirms the model’s ability to detect detector inefficiencies
 138 across different periods.

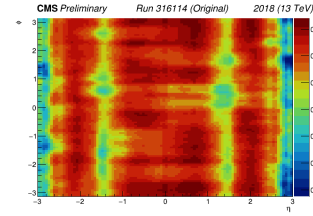


Fig. 7. – Original Hadron Occupancy histogram.

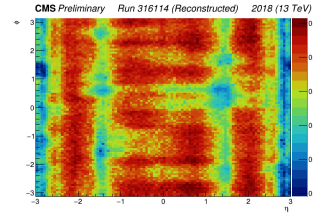


Fig. 8. – Reconstructed histogram.

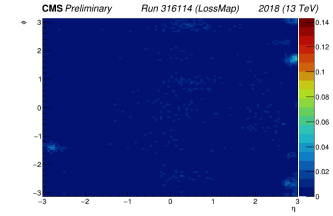


Fig. 9. – MSE loss map.

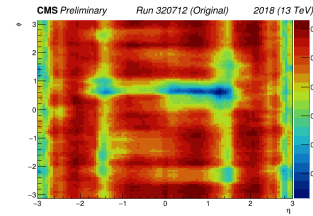


Fig. 10. – Original Hadron Occupancy histogram.

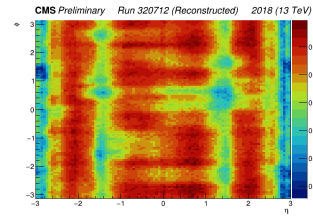


Fig. 11. – Reconstructed histogram.

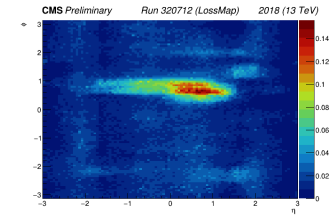


Fig. 12. – MSE loss map.

6. – Conclusion

Unsupervised machine learning methods demonstrated performance comparable to supervised approaches in certifying hadronic jet data. The flexibility of unsupervised learning allows individual jet and MET histograms to be used effectively for anomaly detection:

- Discrepancies in MET distributions can indicate incorrect jet energy measurements in specific detector sectors.
- Training on jet distributions in the Eta-Phi plane reveals variations in jet behavior across Runs.

These results highlight the potential of machine learning to streamline and enhance data certification in the CMS experiment.

REFERENCES

- [1] CMS Collaboration, The CMS experiment at the CERN LHC, *JINST* **3**, S08004 (2008)
- [2] V. Azzolini, B. van Besien, D. Bugelskis, T. Hreus, K. Maeshima, J. Fernandez Menendez, A. Norkus, J. F. Patrick, M. Rovere and M. A. Schneider, EPJ Web Conf. **214**, 02003 (2019), doi:10.1051/epjconf/201921402003
- [3] CMS Collaboration, Machine Learning Techniques for JetMET Data Certification of the CMS Detector, CMS-DP-2023-032