Deep learning for certification of the quality of the data acquired by the CMS Experiment

Adrian Alan Pol[1,3], Virginia Azzolini[1], Gianluca Cerminara[1], Federico De Guio[1,4], Giovanni Franzoni[1], Cecile Germain[3], Maurizio Pierini[1] and Tomasz Krzyżek[1,2] for the CMS Collaboration

[1] CERN, Meyrin, Switzerland
[2] Jagiellonian University, Kraków, Poland
[3] Université Paris-Saclay, Orsay, France
[4] Texas Tech University, Lubbock, Texas, U.S.

E-mail: adrianalan.pol@cern.ch

Abstract. Certifying the data recorded by the Compact Muon Solenoid (CMS) experiment at CERN is a crucial and demanding task as the data is used for publication of physics results. Anomalies caused by detector malfunctioning or sub-optimal data processing are difficult to enumerate a priori and occur rarely, making it difficult to use classical supervised classification. We base out prototype towards the automation of such procedure on a semi-supervised approach using deep autoencoders. We demonstrate the ability of the model to detect anomalies with high accuracy, when compared against the outcome of the fully supervised methods. We show that the model has great interpretability of the results, ascribing the origin of the problems in the data to a specific sub-detector or physics object. Finally, we address the issue of feature dependency on the LHC beam intensity.

1. Introduction

The CMS experiment is one of the two general purpose experiments at the Large Hadron Collider (LHC). The CMS detector is a complex apparatus composed of several sub-detectors, each of them specialized in the measuring the properties of a particular kind of particles. A detailed description of the detector is in [1]. The data acquired by the experiment are scrutinized by a procedure called *data certification* (DC) which ensures they are usable for all physics analysis. This procedure is the last step of the complex Data Quality Monitoring (DQM) [2] apparatus of the experiment. The current certification procedure is conducted by experts of the various sub-detectors and is based on histograms of the relevant quantities which are monitored, at various stages of the data-processing infrastructure, via the DQM setup.

The CMS data, as well as the DQM task, are organized in *acquisition runs*, where duration in time is varying from as little as a few minutes to as much as several hours. Each run is divided into luminosity sections (LSs), an interval corresponding to a fixed number of proton-beam orbits in the LHC and amounting to approximately 23 seconds, numbered progressively from 1 at the start of each run. Each LS can be identified uniquely by specifying the LS number and the run number. In case an anomaly is detected by the certification procedure, the work of pin-pointing the exact times affected by anomalous behavior can require further investigation and the use of non-event data. Besides, when the affected interval of the acquisition run is short, the statistics available in the DQM histograms is often too limited for human assessment. As a consequence, transient problems are difficult to identify and very often a conservative approach has to be adopted discarding more data than actually necessary. The ever increasing physics data volume as well as detector complexity calls for ways to automate this monitoring step: the CMS collaboration is looking into new algorithms allowing it to assess the quality of the physics objects in the reconstructed data with high accuracy and fine time granularity.

The data acquired by the experiment are subdivided into several datasets depending on their physics content. Each dataset undergoes a reconstruction procedure yielding several different collections of physics objects. The certification procedure needs to assess the performance of all of them: this high data dimensionality naturally points toward exploration of deep learning algorithms. Anomalies can be caused by detector malfunctions or sub-optimal software reconstruction and, by nature are rare and not known a priori. Consequently, the use of supervised anomaly detection methods, such as binary classification neural networks, is problematic as a positive (anomalous) class may be misrepresented in the training set. Furthermore, the characteristics of *good* data are evolving with the LHC or CMS configuration. In our research, we base our prototype on a semi-supervised approach which uses deep autoencoders [3], trained on the data acquired during the 2016 LHC campaign.

2. Dataset and Preprocessing

After the data are recorded, the processing step (*reconstruction*) transforms the output of electronic read-outs into human-interpretable variables organized in particle candidates (e.g. photons, muons, jets, tracks, etc.). The Analysis Object Data (AOD) format provides data for physics analysis in a convenient, compact format. It contains a copy of all the high-level physics objects, plus information sufficient to support typical analysis actions. The present study is based on this AOD data format, considered as the best trade-off between the level of reconstruction (number of features needed to describe each LSs) and the amount of information stored in those features. Past research [4] utilized miniAOD dataset that has less features and consequently less information to learn from. The dataset used in the current work consists of all 163684 LSs data recorded from June to October 2016. Several types of reconstructed particle objects are included to maximize the coverage of the algorithms for different physics objects and possible anomalies. This accounts to total of 401 physics variables (e.g. transverse momentum, energy, cluster multiplicity, particle direction). Whenever the ground truth is needed, we rely on the quality labels (*good* or *bad*) determined by a manual certification procedure by detector experts.

As explained earlier, human experts make decisions regarding the data quality based on histograms. When a sub-detector exposes an abnormal behavior e.g. becomes unresponsive, this is reflected in the reconstructed variables. In case of an anomaly, the corresponding histograms should show a considerable deviation from the nominal shape. To mimic the logic of the current procedure we decided to represent each sample as a 2807 dimensional vector that accumulates five quantiles, mean and standard deviation for all 401 variable distributions.

As mentioned before, the physics data is stored in different Primary Datasets (PDs). PDs are subsets of the event stream acquired by the CMS experiment grouped according to the presence of different types of particle candidates. Currently the DC process uses a number of PDs tailored for the physics objective i.e. SingleMuon PD for *muons* or EGamma PD for *electrons*. For our study we have decided to use a dataset defined by the presence of *jets* in reconstructed collision products; the signature of jets involve all of the CMS sub-detectors, enabling our research to be generic and unbound to a specific part of the experiment or type of hardware problem. The proposed strategy has to be generic enough to be applicable for different PDs and in the future it is critical that the performance for all PDs is measured.

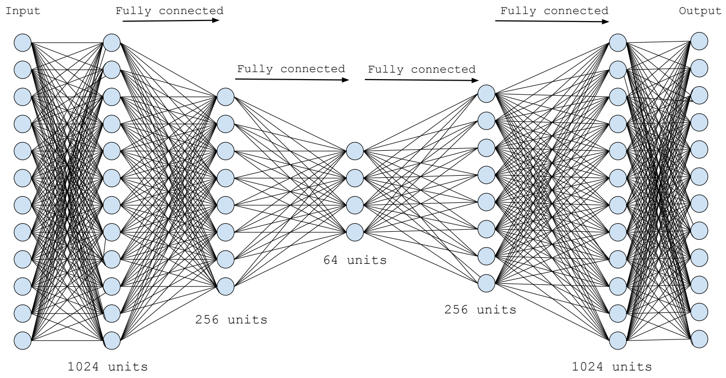

Figure 1. Proposed base architecture. The hyper parameters were chosen using grid search.

3. Methods and Experimental Design

Anomalies are rare in the CMS data: they account for roughly 2% of the dataset which is a small set of examples of failures. Moreover, emerging, unprecedented failures are difficult to anticipate. This makes supervised methods vulnerable to incomplete or inadequate representations of potential failures. A semi-supervised anomaly detection approach models only negative (good) class distribution. During data taking, the model aims at identifying unobserved patterns in newly recorded data. In this manner we intend to retain the full potential to catch all the future and unseen detector failure modes. To this purpose we exploit autoencoders under the assumption that, when trained on a negative class, they yield sub-optimal representations for novel samples. The discrepancy between input and the output indicates that a sample is likely generated by a different (anomalous) process.

We use the autoencoder architecture shown in Figure 1 and propose different regularization techniques [5, 6, 7]. Our sparse autoencoder has $L1$ kernel regularization (10^{-5}) on all of the hidden nodes. The exact penalty term was established using random search. Given the input values are not scaled to a predefined range, we use a parametric rectified linear unit as an activation function in the output layer. Hidden units are also using this type of activation. We train the network with Keras [8] and TensorFlow [9] using the Adam optimizer [10] (with a learning rate of 0.0001, $\beta_1 = 0.7$, $\beta_2 = 0.9$) and early stopping mechanism monitoring the validation dataset with patience set to 32 epochs. The network is instructed to minimize mean squared error (MSE) between input X and the output $\hat{X}$ vector:

$$\epsilon = \frac{1}{n}\sum_{i=0}^{n}(x_i - \hat{x_i})^2$$

Once deployed, the algorithm will evaluate the LSs in the order of recording by the apparatus. To simulate this production scenario, we split the datasets into training (60%), validation (20%) and testing (20%) sets after sorting all samples chronologically. Since both the LHC and the response of the CMS sub-detectors evolve gently with time, random splitting could lead to unintended data snooping [11] where the model is tested on LSs nearly identical to ones in the training set. At early stages of this study it was noticed that the contamination in training sets harms the performance of the algorithm. Thus all the positive samples are removed from training and validation sets. The test set is extended by those anomalous samples previously removed. Including more positive examples in the test is a better approach, as the set has a limited amount of them. This helps qualify the performance of various methods given that bad LSs should always be qualified as bad.

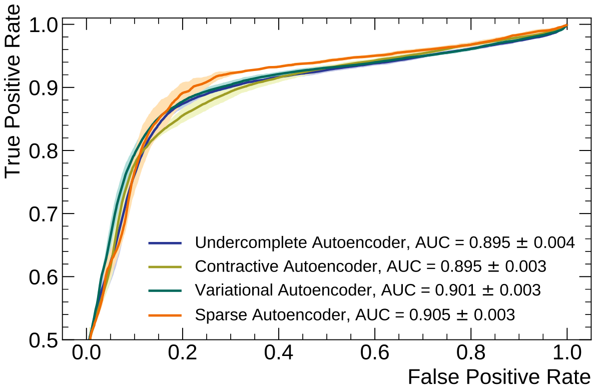

Figure 2. ROC and AUC of different autoencoder models using different regularization techniques.

The difference between reference and recorded distributions is dominated by noise. Experts pay attention only to significant deviations. To mirror this behavior the final decision function is computed using mean squared error of only the worst 100 autoencoder reconstructed features (TOP100):

$$\text{TOP100} = \frac{1}{100}\sum_{i=1}^{100} sorted(x_i - \hat{x}_i)^2.$$

4. Experimental Results and Discussion

The final ROC curves for models and their corresponding AUC are reported in Figure 2. All models show good performance, especially sparse autoencoder.

Figure 3 shows the TOP100 error yield for each sample in the test set as a function of LHC instantaneous luminosity. The instantaneous luminosity is proportional to the number of collisions developing in each bunch crossing multiplied by the number of bunch crossings per second. The error is visibly higher in low and high luminosity regions. This results in the model being unable to capture full data variability. We hypothesized that this dependence was also caused by a smaller amount of samples coming from those regions. Sample weights were used in order to penalize error in those regions more, but no performance improvement was noticed. Adding additional autoencoder input carrying values of instantaneous luminosity has also been shown not to improve the performance.

5. Comparison with Supervised Anomaly Detection

In an evolving conditions context, the CMS Collaboration is looking for tools guaranteeing stable performance over time even at the cost of slightly lower performance. It was expected that the performance of the supervised machine learning algorithms can change dramatically over the course of learning and improve as more data is evaluated and labeled by the experts, and thus available for training. The performance of supervised methods is expected to have some intrinsic limit, especially in periods when novel failures emerge.

We evaluated performance of XGBoost [12] (a supervised method), Isolation Forest and our sparse autoencoder as a function of time. Every 20% of chronologically sorted dataset, which constitutes approximately one month of data taking, all models were retrained using all available past data. Figure 4 shows performance evolution for each model calculated since

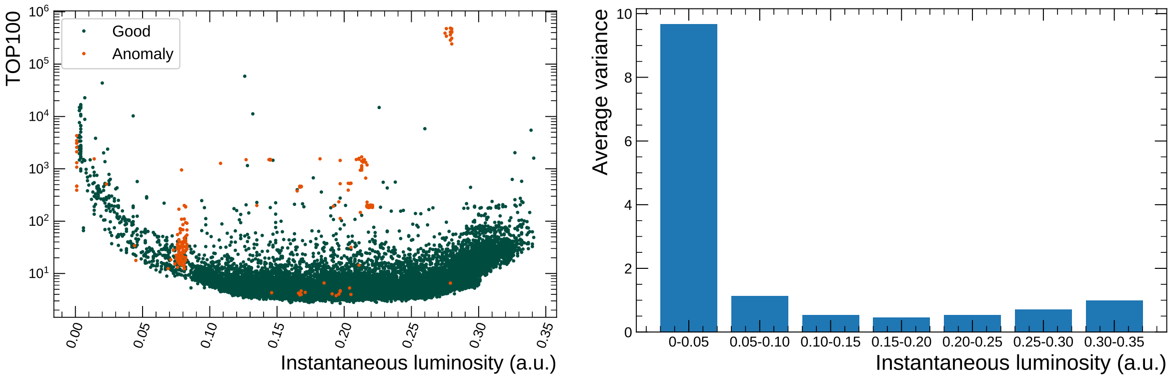

Figure 3. Anomaly score w.r.t. instantaneous luminosity (left) of sparse autoencoder and the average feature spread in different instantaneous luminosity regions (right).

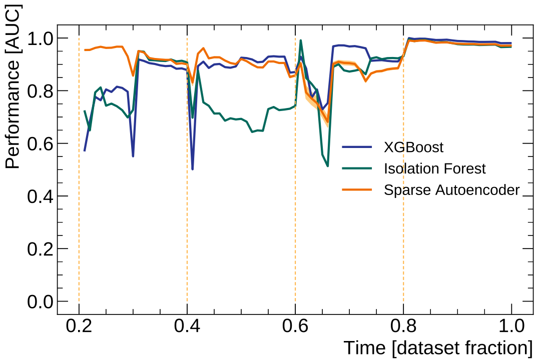

Figure 4. Performance of different strategies as a function of time. After each 0.2 of the dataset, each method is retrained on all past data points. AUC is reported since last retraining.

last retraining time. The visible performance drops, around 0.3, 0.4 and 0.65, are caused by appearance of novel problems. The autoencoder performance is less affected by those events than the performance of XGBoost. Based on those results we conclude that the semi-supervised approach guarantees more stable performance than supervised ones. Nevertheless, a fully supervised approach may still be a powerful addition to the proposed protocol as its performance is frequently superior.

6. Understanding Classification Results

Using the granularity of the MSE, the autoencoder reconstruction can be examined for each feature in a sample. The misbehaving variables whose contribution to the overall error is high can be singled out. This method provides an additional way to interpret the results, in a simple human-consumable form. Figure 5 shows a visualization for such investigation with grouped features (according to different physical meaning). Using human expert knowledge, those plots can map output of a reconstruction to specific detector failures.

7. Conclusions and Outlook

This work explored the usage of autoencoder models for semi-supervised anomaly detection applied to the CMS physics data. The proposed method monitors the distribution of several

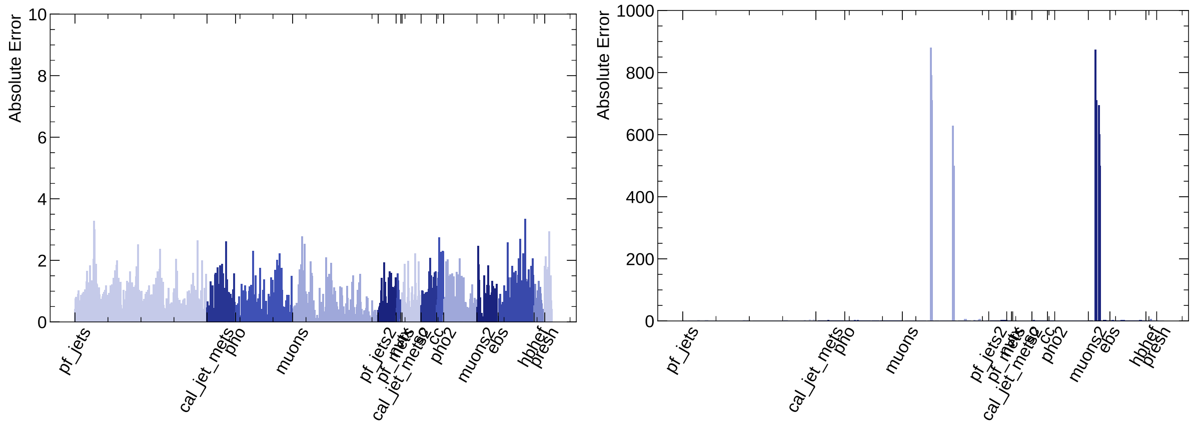

Figure 5. Reconstruction error of each feature for two samples. Different colors represent features linked to different physics objects. For a negative sample (left) we can expect similar autoencoder reconstruction errors across all objects with small absolute scale. Anomalous samples (right) have clearly visible peaks for problematic features (muons).

hundred physics quantities with very fine time granularity, allowing to identify emerging anomalies promptly and to clearly identify which ones among the input variables show an anomalous behavior. This aspect of the interpretability of the results is a key feature for the physicists operating the tool. The model proved robust against rare and newly emerging anomalies in the available dataset. Further studies are needed to consolidate a training and deployment strategy allowing the model to accurately describe the evolving nature of the experiment data.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement n^{o} 772369).

References

[1] S. Chatrchyan et al. (CMS), *The CMS experiment at the CERN LHC* (2008). JINST **3**

[2] M. Schneider, *The Data Quality Monitoring Software for the CMS experiment at the LHC: past, present and future*, in *Proceedings to CHEP 2018* (2018)

[3] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning (pages 499-523)* (MIT Press, 2016)

[4] M. Borisyak, F. Ratnikov, D. Derkach, A. Ustyuzhanin, *Towards automation of data quality system for CERN CMS experiment*, in *IOP Conf. Ser J Phys Confer Ser* (2017, doi: 10.1088/1742-6596/898/9/092041), **898**, p. 092041

[5] M. Ranzato, C. Poultney, S. Chopra, Y. LeCun, *Efficient Learning of Sparse Representations with an Energy-based Model*, in *Proceedings of NIPS* (2006), pp. 1137–1144

[6] D.J. Rezende, S. Mohamed, D. Wierstra, *Stochastic Backpropagation and Approximate Inference in Deep Generative Models*, in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32* (2014), ICML'14, pp. II–1278–II–1286

[7] S. Rifai, P. Vincent, X. Muller, X. Glorot, Y. Bengio, *Contractive auto-encoders: Explicit invariance during feature extraction*, in *Proceedings of the 28th international conference on machine learning (ICML-11)* (2011), pp. 833–840

[8] F. Chollet et al., *Keras*, https://keras.io (2015)

[9] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard et al., *Tensorflow: a system for large-scale machine learning.*, in *OSDI* (2016), **16**, pp. 265–283

[10] D. Kingma, J. Ba, *Adam: A method for stochastic optimization* (2014). arXiv:1412.6980

[11] H. White, *A reality check for data snooping* (2000). Econometrica **68**

[12] T. Chen, C. Guestrin, *Xgboost: A scalable tree boosting system*, in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (ACM, 2016), pp. 785–794