PROCEEDINGS OF SCIENCE

Using machine learning techniques for Data Quality Monitoring in CMS and ALICE

Kamil Deja* **for the CMS and ALICE Collaborations**
Institute of Computer Science, Warsaw University of Technology
E-mail: K.Deja@ii.pw.edu.pl

Data Quality Assurance plays an important role in all high-energy physics experiments. Currently used methods rely heavily on manual labour and human expert judgements. Hence, multiple attempts are being undertaken to develop automatic solutions especially based on machine learning techniques as the core part of Data Quality Monitoring systems. However, anomalies caused by detector malfunctioning or sub–optimal data processing are difficult to enumerate a priori and occur rarely, making it difficult to use supervised classification. Therefore, researchers from different experiments including ALICE and CMS work extensively on semi–supervised and unsupervised algorithms in order to distinguish potential outliers without manually assigned labels. In this contribution, we will discuss several projects whose that aim at solve this task. Machine learning based solutions bring several advantages and may provide fast and reliable data quality assurance, simultaneously reducing the manpower requirements. A good example of this approach is a model based on deep autoencoder employed in the CMS experiment which has been successfully qualified on CMS data collected during the 2016 LHC run . Tests indicate that this solution is able to detect anomalies with high accuracy and low fake rate when compared against the outcome of the manual labelling by experts.

Researchers from the ALICE experiment are currently working on a similar task. They intend to perform a data quality checks in much higher granularity. The current approach is limited to run classification based on manually set cut–offs on descriptive data statistics. More sophisticated machine learning based methods may enable more accurate data selection, on high granularity level of 15-minutes data acquisition periods.

7th Annual Conference on Large Hadron Collider Physics - LHCP2019
20-25 May, 2019
Puebla, Mexico

*Speaker.

PoS(LHCP2019)236

1. Introduction

The goal of all High Energy Physics experiments is to provide accurate and real measurements of physical phenomena. This is why, examination of the quality of gathered data is a crucial part of their employment. In both A Large Ion Collider Experiment (ALICE) [1] and Compact Muon Solenoid (CMS) [2] this process consists of two important parts: Online Data Quality Monitoring, and Offline Data Quality Assurance.

Data Quality Monitoring (DQM) is an important aspect of every High-Energy Physics experiment, especially in the era of the Large Hadron Collider (LHC) [3] where detectors are extremely sophisticated and delicate devices. To avoid recording low quality data, one needs an online feedback on the quality of the data being recorded for offline analysis. DQM software provides this feedback and helps shifters and experts to identify early potential issues. DQM involves the online gathering of data, their analysis by user-defined algorithms and the storage and visualisation of the produced monitoring information [4, 5]. In the ALICE experiment, during the past data taking periods (Run 1, Run 2), more than 60 DQM agents have been continuously running in production, monitoring the incoming 6 GB/s of data and producing around 10000 objects, mostly 1D and 2D histograms, updated every minute or more.

This very large amount of information is mostly intended for experts. An operator, sitting 24/7 in the ALICE Run Control Room, is checking a small subset (< 100) of these histograms using information provided by the experts via visual clues and textual documentation. A number of simple checks are also done automatically, such as having the mean within certain limits defined manually by the detectors specialists. The process is therefore very much human-based and leads to massive amount of tedious labour being carried out by people. At the same time, the second part of processing – Data Quality Assurance (DQA) systems currently employed in HEP experiments are also based on manual, human-based comparisons of data statistical distributions and their parameters.

In this context, machine learning revolutionises the way in which data quality is monitored in the High Energy Physics experiments. With the increase of system complexity and data throughput, it becomes less and less conceivable to manually evaluate the quality of the collected data. Machine learning is starting to be used to identify patterns in the stream of data and the statistical preprocessed variables to raise the alarms when detecting anomalies or to tag abnormal chunks of data which should not be used for analyses.

2. Machine learning methods for quality assurance

Machine learning based quality assurance methods have already proven their usefulness in various domains including computer security [6, 7], medicine [8] or space technology [9]. Based on the nature of the collected data we can distinguish two approaches into the problem: supervised and unsupervised learning.

3. Supervised quality assurance methods

If only ground truth information about the exemplar training data quality is present (e.g. label assigned by an expert to each example) we can benefit from it using supervised machine learning

methods (e.g. [10]). Their goal is to search for patterns in data which may indicate whether or not a given example should be treated as anomaly according to training examples and their labels. Usually methods used for this task are optimised to find the most adequate border which distinguish those examples marked originally as "good" from those labelled as "bad".

Model trained on the subset of available data, can be then generalised on the whole domain. This task known as classification is a common approach which can be addressed using numerous different algorithms. Starting from the simplest ones e.g. Bayesian classification, SVM [11] or Decision Trees, ending up in the current state of the art methods based on Deep Neural Networks and Boosted decision trees [12].

Supervised quality measurement methods are widely used in practical Data Science approaches. However, these techniques can also be beneficial for certain tasks in High Energy Physics experiments. As already mentioned, in both ALICE and CMS experiments, there are parts of data processing which require manual assessment of data quality by highly qualified detector experts. This is why, there are several projects which aim to develop predictive model based on the historical expert decisions collected through years of data acquisition. Such models may be then used to seek for similar patterns in new data, which may indicate most probable decision of real expert. In this paper, we describe one of the task performed in the CMS experiment which benefit from such approach.

4. Unsupervised quality assurance methods

While classification of data examples quality provides accurate results, it is often the case that for many problems we cannot easily collect ground truth labels of training data. In such situation we still may use machine learning techniques to seek for a examples which are the most abnormal. There are several different methods to define outlying observations. Accordingly, based on those observations various algorithms may be used in order to find data anomalies. Notable approaches use statistics [13] or clustering [14] in order to find observations which are far from modelled profile of normal instances. With an advancements in machine learning, novel algorithms seek for anomalies in a different way.

Isolation Forest [15] is a technique based on the random forest [16] which is a robust classification algorithm. In the standard random forest the goal is to properly distinguish examples from different classes by creating numerous decision trees. The idea behind isolation forest extends this processing with a simple observation. Supposing that each example is a singe representative of a different class, we may need different number of decision trees (cut-offs) of a random forest algorithm, to distinguish it from all of the other classes. If the given example exists in the highly populated area probably we would need more decision trees to isolate it from surrounding neighbours (see x_i in Fig. 1). At the same time, observations in sparse areas, or at the edge of the distribution are easily separable. Therefore, isolation forest builds ensemble of so called iTrees and select as anomalies those observations which have the shortest average path length through the trees. This processing allows not only to select observations which are abnormal, but also to rank all of the processed data examples in terms of their abnormality.

Anomaly detection with autoencoders

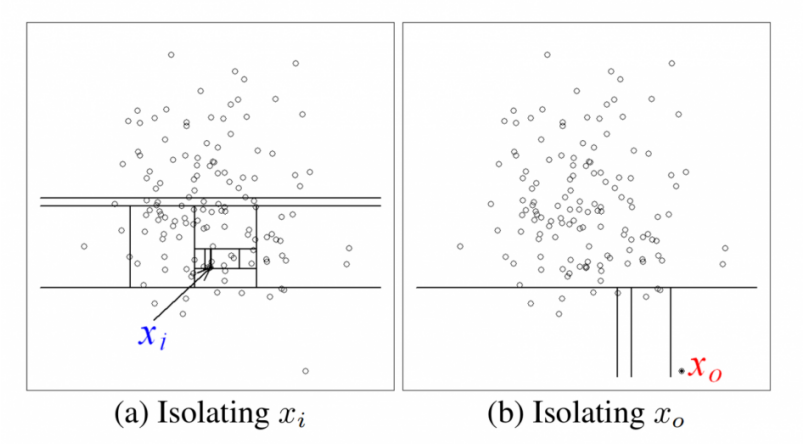


Figure 1: Visualisation of the principle idea behind Isolation forest. Anomalies (x_0) are easiest to separate from surrounding neighbours, while a normal point (x_i) requires twelve random partitions to be isolated. Visualisation from [15].

Autoencoders as an artificial neural networks model are usually employed in terms of data dimensionality reduction and data compression [17, 18]. However, thanks to their generalisation performance, they can be also beneficial for unsupervised and semi-supervised anomaly detection.

As observed in [19, 20], abnormal observations are usually underrepresented in the latent space of properly trained autoencoder. This observation may be easily explained by the fact that it is profitable (in terms of network's loss function), to make mistakes in the rarest types of observations.

On the basis of this observation, we can train autoencoder on the historical training data and rank the data examples by the autoencoder's average reconstruction error. If only possible, in order to emphasise the difference between good quality examples and anomalies, we can perform a semi-supervised learning using only the part of a training data which was originally labelled as "good". In such a case network fits to a good quality data examples what results in even higher reconstruction errors for anomalies. Experiments with this approach have already been conducted in both CMS and ALICE experiments.

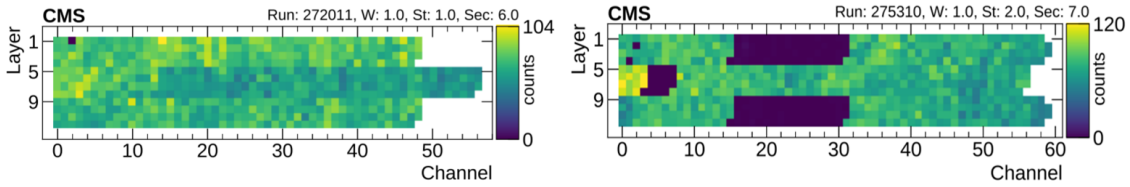
5. Current development of machine learning methods for Quality Assurance in CMS and ALICE

Although majority of the quality monitoring systems in High Energy Physics experiments currently employed at LHC is based on standard, manual approach, there are several attempts to employ machine learning methods described above. In this section we will discuss two use-cases from the CMS experiment and one from ALICE.

5.1 Classification of drift tube hit occupancy in CMS

Problem of the drift tube hit occupancy is a perfect example of situation where qualified experts are obliged to monitor data quality, in this case represented by the hit occupancy histograms. Those histograms contains the total number of electronic hits at each readout channel presented as a 2-dimensional array organised along layer (row) and channel (column) indices. At any time there is

a total of 250 histograms visualised by the CMS DQM tool. Previous processing, based entirely on the human expert judgement required 24-hour monitoring of hit occupancy histogram in the control room. To help in anomalies identification there was an automatic alarming system triggered when the fraction of dead cells (zero hits) in one of the plots was above a given threshold. This is visible in the Figure 2(b) which presents real data example of low occupancy across all 12 layers.



(a) (b)

Figure 2: Input data examples: **(a)** of the good quality data and **(b)** bad quality data with visible low occupancy regions.From [21]

To address the shortcoming of this processing, researchers from CMS experiment proposed a machine learning solution based on three strategies which search for ineffective areas of drift tubes at different scales [22]:

- **Local** – independent anomalies selection for data collected in each chamber layer
- **Regional** – extension of the local approach to account for intra-chamber problems. Simultaneously consider all layers in a chamber, but each chamber independently from the others
- **Global** – use the information of all the chambers for a given acquisition run.

For the purpose of the studies, authors performed tests using 6280 data examples annotated by detector experts. In this dataset 612 samples (around 10%) were denoted as bad quality. After preprocessing which included data standardisation, smoothing and normalisation authors performed a supervised classification task.

For the local anomaly search strategy, the best performing method is based on deep convolutional neural networks [23]. The exact architecture consisted of three convolutional layers with rectified linear units and two fully connected layers with softmax as activation functions. Network was trained using Keras/TensorFlow [24] and Adam [25] optimiser. The performance of this approach reached almost the quality of detector experts with AUC equal to 0.995. Therefore, this part of solution presented in [22] was already successfully implemented in the CMS DQM monitoring system.

For both next steps regional and global anomaly search, the best performing solution was based on the semi-supervised convolutional autoencoder. Therefore, for the purpose of this experiments, researchers extended the training dataset using all of the examples not qualified by the experts as bad quality. However, the results of this approaches was not yet as satisfying reaching AUC=0.944.

5.2 Semi-supervised anomaly detection for offline quality assurance in CMS

On the other hand, current offline quality assurance in the CMS experiment is also based on the extensive labour of human experts who classify data quality on a run by run basis. With the

ongoing improvements of LHC, this solution might be no longer suitable. Therefore in [26, 21], researchers from the CMS collaboration proposed a novel solution for this problem based on the machine learning approach with autoencoders. Additionally, to ensure the biggest possible sample of valuable data for physical studies, authors propose to assess data quality on a per-luminosity section (i.e. 23 seconds of data taking) basis.

For the purpose of the studies, authors gathered data samples from June to October 2016 data taking periods. Each of data samples consisted of 401 variables (eg. energy, eta, phi etc.). For each of variables, authors calculated its five quantiles, mean and rms values which resulted in 160 000 data samples with the total of 2807 features.

The main problem of data quality assessment with machine learning methods is the sparsity of anomalies. In the case of CMS offline monitoring, it is around 2% of anomalies in the prepared dataset. This class imbalance makes it extremely hard for supervised methods to capture patterns responsible for abnormal readings. Therefore, authors decided to use semi-supervised method based on autoencoders to learn the distribution of good quality samples, hoping to reveal all unseen detector failures.

To find the best performing model, authors run experiments on several approaches. Namely: standard autoencoder, sparse autoencoder(L1,(10^-5)) [27],contractive autoencoder [28] and variational autoencoder [29]. For all of the approaches a deep architecture was trained, based on 5 fully connected hidden layers, with 64 neurons in latent space. Models were trained to minimise MSE between input and output layers, using Adam [25] optimiser. Dataset was divided into train and test subsets. For the first one the initial 80% of all good quality examples was used (in chronologically sorted order). The second one consisted of remaining 20% of good quality data as well as all anomalies.

Studies reviled that sparse autoencoder architecture performed the best in terms of anomalies discovery reaching AUC = 0.905 $\pm$ 0.003. Exact results of the studies are presented in Figure Fig. 3.

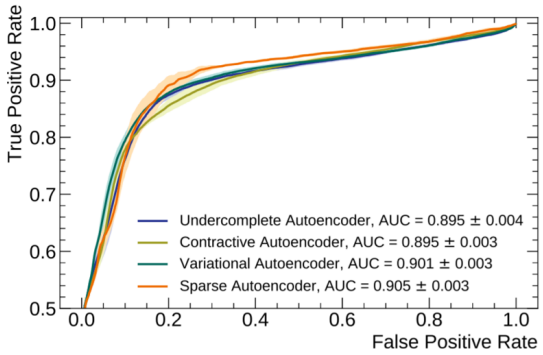
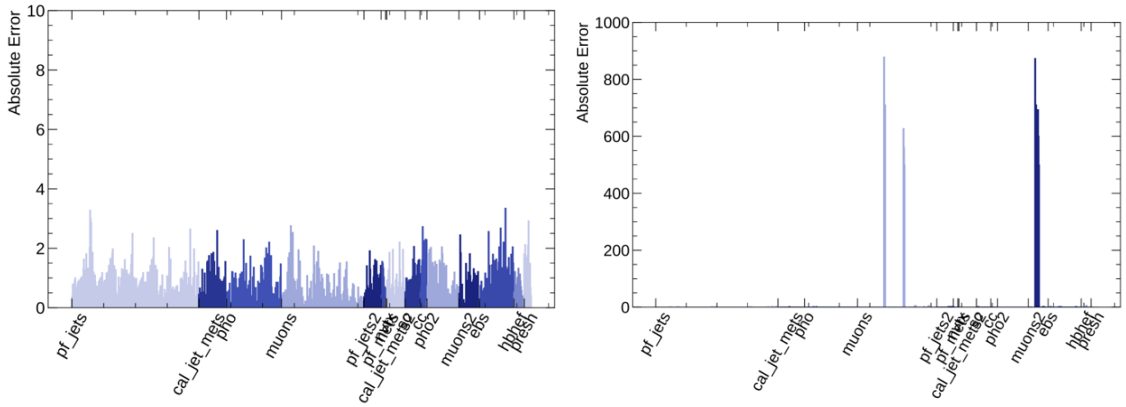


Figure 3: Results of the automatic anomaly detection with autoencoders: ROC plot of different autoencoder methods. From [26]

What is also of the great interest for quality assurance experts, results obtained with autoencoders are easy to interpret. As presented in Fig. 4 good quality examples as Fig. 4(a) have by default low absolute reconstruction error on all parameters. On the other hand, for bad quality ex-

ample (Fig. 4(b)) we can see very high reconstruction error on certain problematic attributes with abnormal values.



(a) (b)

Figure 4: Examples of the reconstruction error for all data attributes for good **(a)** and bad **(a)** quality sample. From [26]

5.3 Semi-supervised anomaly detection for offline quality assurance in ALICE

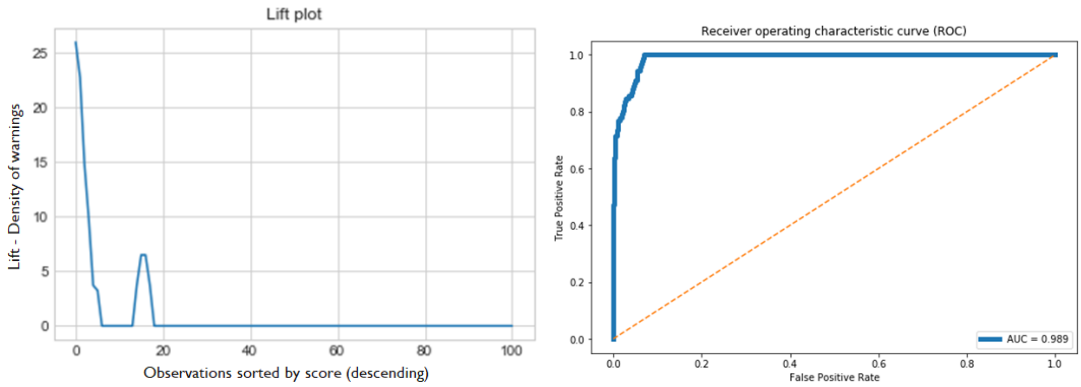
Based on this research conducted in the CMS experiment, we propose a similar solution for ALICE. To this end, we present a preliminary results of experiment performed on the Time Projection Chamber (TPC) which is the main tracking detector of ALICE.

Current data quality assurance in ALICE is based on the automatic system, which was in use during Run2 data collection period. This system monitored around 40 parameters defined by detector experts. Some of these parameters were for example related to low-level features of the reconstructed tracks (eg. distance of closest approch to primary vertex - DCA) or gathered information from all reconstructed tracks (e.g. track multiplicity). Similar to CMS, quality assurance in ALICE was performed run-wise. In order to assign a data quality label, for each period of data gathering (usually dozens of runs) the robust mean and standard deviation of all 40 parameters were calculated, using 80% of the data which minimises the variance. Based on those global mean and standard deviation, each run was assign a quality label in such way, that all runs for which mean value of any parameter deviated from the global mean by more than 3/5 σ, were classified as warnings/outliers. This very simple processing provides high purity of data, but at the same time rather limited efficiency. Therefore, we propose a machine learning based solution with autoencoders which aims in automatic data quality evaluation.

For the purpose of machine learning experiments, the set of tracked parameters was extended (to over 200) by quantities related to working conditions inside the detector. Additionally, we divided the data acquisition period into short 10-15 min chunks in order to provide high granularity data assessment. We gathered the data from five acquisition periods, two Pb-Pb (LHC18q,LHC18r), and three p-p (LHC18f,LHC18o,LHC18p). Then we performed a standard data assurance procedure to assign to each data chunk an automatic quality label. In total this processing resulted in 2508 data examples out of which 91 were identified as warnings and 71 as outliers.

At first, to ensure that we can expect the same, high fidelity results with machine learning solution as with standard methods, we performed a supervised learning task using xgboost algorithm [12]. We trained the algorithm to predict automatically assigned warnings. The results of this approach were satisfying, reaching precision and recall on test-set equal to 98%. Therefore, further studies are performed with unsupervised learning methods to seek for other ways to determine which of the data examples should be classified as abnormal.

We start with the Isolation Forest algorithm which arrange data examples by assigning them abnormality score. Although, the method works in entirely unsupervised way, we confront these scores with automatically assigned quality labels. As presented in the Fig. 5(a), we can observe a high concentration of anomalies in the first few percentiles of data examples with highest score. This means that isolation forest treats as anomalies mostly the same examples as a standard method. However, the maximum at around 18% suggests the number of observations which were automatically classified as outliers, while at the same time, they were assigned low score from isolation forest. With closer examination we discovered that these observations only slightly exceed the threshold of 3 σ distance from global mean value on several parameters. Therefore, it is possible that these samples should not be treated as anomalies for certain physical studies.

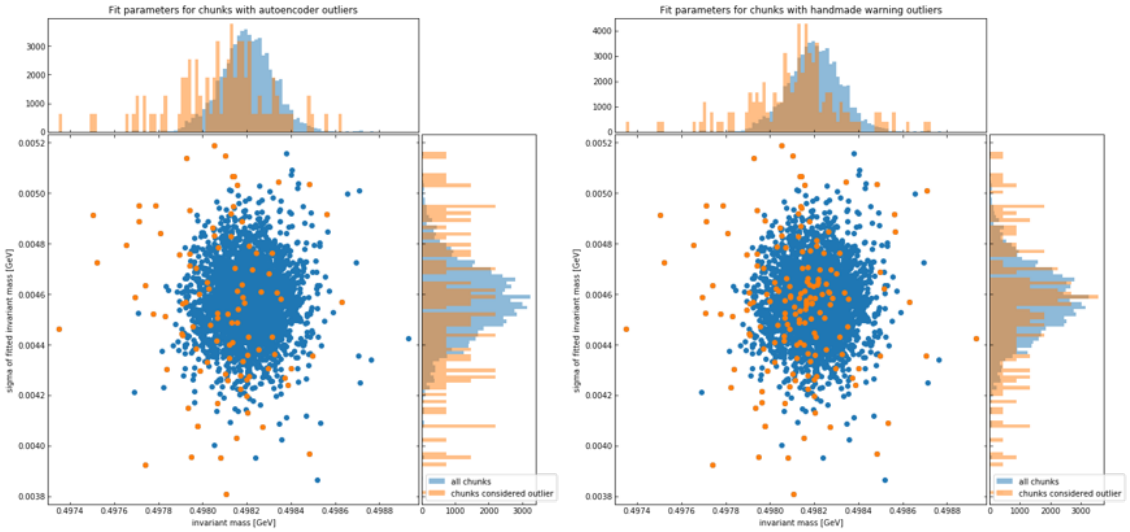


(a) (b)

Figure 5: Lift score of data chunks scored by Isolation Forest comparing to automatically assigned labels.**(a)**, and ROC curve plot for semi-supervised anomalies detection with autoencoder.**(b)**

To analyse further the possibilities with unsupervised learning we propose a solution similar to the one introduced in CMS experiment. Therefore we create a deep autoencoder to provide more accurate and explainable solution. We propose a bottleneck autoencoder architecture with five hidden layers. We train the network in semi-supervised method as proposed in [21] using 75% of good quality examples. We then once more confronted the results with standard quality assurance procedures. As visible in the ROC plot (Fig. 5(b)), with fine tuned training, we are able to recreate almost the same classification as with the standard procedure. On the other hand, we can tune an autoencoder to provide softer cuts on data and therefore increase the efficiency of data selection. Nevertheless, additional quality checks are needed to ensure that the purity of newly selected additional data samples is preserved.

To that end we perform additional studies, by running simple physical analysis on the whole collected data examples. For each data chunk we fit a peak of invariant mass for K^0_S. As presented in Fig. 6, our approach with autoencoders tend to tag as outliers mostly data chunks for which the



(a) (b)

Figure 6: Mean and σ of K^0_S invariant mass for outliers selection with autoencoders **(a)**, and standard procedure **(b)**.

invariant mass lies on far edges of the global distribution (Fig. 6(a)). On the contrary, standard anomalies detection method selects also data examples with central mass selection (Fig. 6(b)).

6. Summary

Machine learning methods are the future for quality monitoring and controlling systems in High Energy Physics. With upcoming upgrades of LHC and experiments themselves in Run3, current processing techniques which are highly dependent on manual labour will not be sufficient. Therefore, future development of reliable and explainable Machine learning solutions is crucial. Recent attempts to tackle different offline and online processes in CMS and ALICE experiments are promising and may be a good baseline for redesigned DQM systems.

7. Acknowledgements

The authors acknowledge the support from the Polish National Science Centre grants no. UMO-2016/21/D/ST6/01946, UMO-2016/22/M/ST2/00176, UMO-2018/31/N/ST6/02374 and the Ministry of Science and Higher Education grant no. DIR/WK/2016/2018/17-1.

References

[1] K. Aamodt et al. “The ALICE experiment at the CERN LHC”. In: *JINST* 3 (2008), S08002. DOI: 10.1088/1748-0221/3/08/S08002.

[2] Albert M Sirunyan et al. “Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13TeV$”. In: (2018).

[3] Lyndon Evans and Philip Bryant. "LHC Machine". In: *JINST* 3 (2008), S08001. DOI: 10.1088/1748-0221/3/08/S08001.

[4] Adriana Telesca et al. "The alice data quality monitoring system". In: *2010 17th IEEE-NPSS Real Time Conference*. IEEE. 2010, pp. 1–6.

[5] Barthélémy von Haller et al. "The ALICE Data Quality Monitoring: qualitative and quantitative review of three years of operations". In: *Journal of Physics: Conference Series*. Vol. 513. 1. IOP Publishing. 2014, p. 012038.

[6] Terran Lane and Carla E Brodley. "An application of machine learning to anomaly detection". In: *Proceedings of the 20th National Information Systems Security Conference*. Vol. 377. Baltimore, USA. 1997, pp. 366–380.

[7] Vipin Kumar. "Parallel and distributed computing for cybersecurity". In: *IEEE Distributed Systems Online* 10 (2005), p. 1.

[8] Clay Spence, Lucas Parra, and Paul Sajda. "Detection, synthesis and compression in mammographic image analysis with a hierarchical image probability model". In: *Proceedings IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA 2001)*. IEEE. 2001, pp. 3–10.

[9] Ryohei Fujimaki, Takehisa Yairi, and Kazuo Machida. "An approach to spacecraft anomaly detection problem using kernel feature space". In: *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM. 2005, pp. 401–410.

[10] Naoki Abe, Bianca Zadrozny, and John Langford. "Outlier detection by active learning". In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2006, pp. 504–509.

[11] Johan AK Suykens and Joos Vandewalle. "Least squares support vector machine classifiers". In: *Neural processing letters* 9.3 (1999), pp. 293–300.

[12] Tianqi Chen and Carlos 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.

[13] Peter J Rousseeuw and Katrien Van Driessen. "A fast algorithm for the minimum covariance determinant estimator". In: *Technometrics* 41.3 (1999), pp. 212–223.

[14] Zengyou He, Xiaofei Xu, and Shengchun Deng. "Discovering cluster-based local outliers". In: *Pattern Recognition Letters* 24.9-10 (2003), pp. 1641–1650.

[15] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation forest". In: *2008 Eighth IEEE International Conference on Data Mining*. IEEE. 2008, pp. 413–422.

[16] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[17] Jürgen Schmidhuber. "Deep learning in neural networks: An overview". In: *Neural networks* 61 (2015), pp. 85–117.

[18] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[19] Mayu Sakurada and Takehisa Yairi. "Anomaly detection using autoencoders with nonlinear dimensionality reduction". In: *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*. ACM. 2014, p. 4.

[20] Guillaume Alain and Yoshua Bengio. "What regularized auto-encoders learn from the data-generating distribution". In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 3563–3593.

[21] Adrian Alan Pol et al. "Detector monitoring with artificial neural networks at the CMS experiment at the CERN Large Hadron Collider". In: *Computing and Software for Big Science* 3.1 (2019), p. 3.

[22] Adrian Alan Pol et al. *Online detector monitoring using AI: challenges, prototypes and performance evaluation for automation of online quality monitoring of the CMS experiment exploiting machine learning algorithms.* Tech. rep. 2018.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[24] François Chollet et al. *Keras*. https://github.com/fchollet/keras. 2015.

[25] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *CoRR* abs/1412.6980 (2014).

[26] Adrian Alan Pol. *Anomaly detection using Deep Autoencoders for the assessment of the quality of the data acquired by the CMS experiment*. Tech. rep. 2018.

[27] Christopher Poultney, Sumit Chopra, Yann L Cun, et al. "Efficient learning of sparse representations with an energy-based model". In: *Advances in neural information processing systems*. 2007, pp. 1137–1144.

[28] Salah Rifai et al. "Contractive auto-encoders: Explicit invariance during feature extraction". In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. Omnipress. 2011, pp. 833–840.

[29] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *CoRR* abs/1312.6114 (2013).