

Real-time Anomaly Detection at the L1 Trigger of CMS Experiment

Abhijith Gandrakota^{a,*} for the CMS collaboration

^a*Fermi National Accelerator Laboratory,
Batavia, IL, USA*

E-mail: abhijith.gandrakota@cern.ch

We present the preparation, deployment, and testing of an autoencoder trained for unbiased detection of new physics signatures in the CMS experiment Global Trigger (GT) test crate FPGAs during LHC Run 3. The GT makes the final decision whether to readout or discard the data from each LHC collision, which occur at a rate of 40 MHz, within a 50 ns latency. The Neural Network makes a prediction for each event within these constraints, which can be used to select anomalous events for further analysis. The GT test crate is a copy of the main GT system, receiving the same input data, but whose output is not used to trigger the readout of CMS, providing a platform for thorough testing of new trigger algorithms on live data, but without interrupting data taking. We describe the methodology to achieve ultra low latency anomaly detection, and present the integration of the DNN into the GT test crate, as well as the monitoring, testing, and validation of the algorithm during proton collisions.

*42nd International Conference on High Energy Physics - ICHEP 2024
18-24 July 2024
Prague, Czech Republic*

*Speaker

1. Introduction

The Large Hadron Collider (LHC), operating at the energy frontier, collides protons at an unprecedented rate of 40 million times per second (40 MHz). The Compact Muon Solenoid (CMS) experiment studies these collisions to uncover potential Beyond Standard Model (BSM) physics and precisely measure rare Standard Model (SM) processes [2]. While the high collision rate at the LHC increases the probability of producing and detecting rare processes, the nearly 100 million channels of the CMS detector also generate an enormous amount of data [10, 14].

Only a small fraction of the 40 MHz proton-proton collision events—around 1,000 per second—can be stored for detailed offline analysis. To meet this stringent data reduction, events are selected using a two-tiered trigger system. The first level (L1), composed of custom hardware processors built with field-programmable gate arrays (FPGAs), uses information from the calorimeters and muon detectors to select events at a rate of around 100 kHz within a fixed latency of $4\mu\text{s}$ [14]. The second level, the high-level trigger (HLT), consists of a processor farm running optimized event reconstruction software, reducing the rate to around 1 kHz before storage [10]. The L1 trigger discards the largest fraction of events, playing a crucial role in the two-tier trigger system. While this approach has enabled the discovery of the Higgs boson and many SM measurements, it risks missing unexpected new physics signatures not anticipated by predefined selection criteria.

This reliance on specific criteria introduces the risk of bias, potentially limiting the ability to detect new physics in the absence of strong theoretical predictions. This highlights the need for more general, unbiased triggering methods that maintain sensitivity to a wide range of new physics signatures. Ideally, such methods should identify anomalous events that deviate from expected SM patterns without relying heavily on specific BSM model predictions. For this reason, there are ongoing efforts to explore unsupervised machine learning (ML) techniques, such as anomaly detection (AD) methods, to select events with potential BSM signatures [7].

In this work, we present a novel approach: the implementation of ML-based real-time AD algorithms at the CMS L1 Trigger level. This cutting-edge technique provides a model-independent method for identifying potentially interesting events while maintaining high rejection rates for background processes. We discuss the development, implementation, and preliminary results of this approach, recently deployed at the CMS L1 Trigger [3]. By introducing advanced ML capabilities at the earliest stage of event selection, CMS is entering a new era of data collection strategy, enhancing the search for unknown phenomena and potentially revolutionizing the approach to triggering in high-energy physics experiments.

2. Anomaly detection L1 triggers in CMS

The CMS collaboration has developed two complementary approaches to anomaly detection at the L1 trigger: AXOL1TL (Anomaly eXtraction Online Level-1 Trigger aLgorithm) and CI-CADA (Calorimeter Image Convolutional Anomaly Detection Algorithm). Both methods leverage advanced machine learning techniques to identify atypical events in real-time, offering a model-independent approach to capturing potential new physics signatures.

One of the well-known ML techniques for anomaly detection is the use of autoencoder (AE) [7]. Autoencoders are neural networks designed to learn efficient encodings of the dominant SM

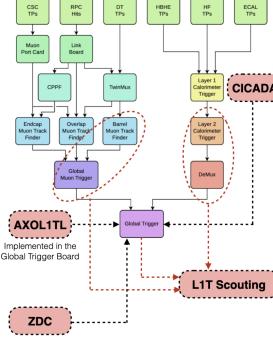


Figure 1: Components of the L1 trigger, along with the input and output paths for the anomaly detection algorithms.

process in the input data. The architecture of an autoencoder consists of two main components: an encoder, which compresses the input data into a lower-dimensional representation (often called the latent space), and a decoder, which attempts to reconstruct the original input from this compressed representation.

These neural networks(NN) are trained to minimize the difference between the input and the reconstructed output. When trained on the proton-proton collisions as detected by the CMS detector, the network efficiently reconstructs the dominant background processes. When presented with an event containing processes BSM physics or rare SM physics; an anomalous event, the autoencoder struggles to reconstruct it accurately. This reconstruction error serves as our anomaly score; events with high reconstruction errors are flagged as potentially interesting.

Both the anomaly detection algorithms are trained on the trigger less stream, (ZeroBias dataset) collected by the CMS experiment in 2023 at a center-of-mass energy of $\sqrt{s} = 13.6$ TeV, where lumineveling at pile-up 62 was used. 50% of this used for training and 50% was used for validation, performance testing. The anomaly detection triggers are estimated to have better performance than a traditional rule based trigger with similar event collection rate. In the case of AXOL1TL, it is to have 46% efficiency gain when compared to rest of L1 trigger, when operating at a rate of $1kHz$ for capturing exotic decay of higgs to four b quarks.

2.1 Inputs to L1 anomaly detection

AXOL1TL and CICADA use different L1 reconstructions as inputs. AXOL1TL takes in 10 jets, 4 electron/photon objects, 4 muons, and transverse missing energy (MET) as reconstructed in the L1 trigger from calorimeter and muon triggers. The 3-momenta (p_T, η, ϕ) of these objects, in raw hardware integer values, are used. CICADA, by contrast, uses calorimeter region energies, which form an 18×14 image-like input. The input streams for both algorithms are shown in Fig. 1.

2.2 Network Architecture and FPGA implementations

AXOL1TL uses a variation of autoencoders called Variational Autoencoders (VAEs), which impose additional constraints on the latent space [11]. The encoder computes a latent space vector of Gaussian distributions $N(\mu = 0, \sigma = 1)$, with a latent vector size of 8. VAEs minimize both the reconstruction error and the deviation from a standard normal distribution by penalizing the KL

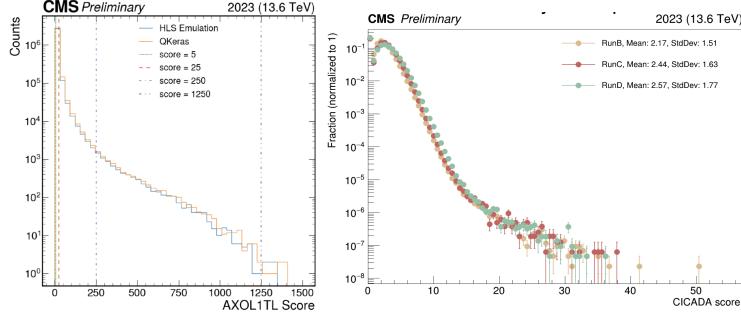


Figure 2: Anomaly score distributions from AXOL1TL (left) and CICADA (right) as output by both HLS emulation and qkeras. Higher-scoring events are flagged as anomalous for further analysis.

divergence (\mathcal{D}_{KL}). The encoder and decoder are dense feed-forward networks. Only the encoder is used for real-time inference, and the anomaly score is approximated by the sum of squared means of the latent vector ($\sum_{i=1}^8 \mu_i^2$).

CICADA uses convolutional layers in its autoencoder [12], as it processes image-like inputs of 252 (18x14) transverse energy deposits in two channels. The reconstruction quality is measured by Mean Squared Error (MSE) loss. To enable ultra-low latency inference, the autoencoder is shrunk using knowledge distillation to mimic the MSE loss with the given input [9, 13].

Both algorithms are trained using quantization-aware training frameworks like qkeras [4]. The networks are then translated to high-level synthesis (HLS) code using `hls4ml` [1, 5, 6], and deployed on Xilinx Virtex-7 FPGAs for ultra-fast inference. These FPGA implementations are carefully tuned for bit-exactness, as shown in Fig. 2.

3. Performance of AXOL1TL

Out of both anomaly detection triggers, AXOL1TL was commissioned for real-time testing and data-taking with the CMS trigger framework. The AXOL1TL trigger was tested at five different thresholds for triggering anomalous events in the data: very tight, tight, nominal, loose, and very loose. Events triggered through the nominal seed are sent for further reconstruction with HLT scouting and stored with minimal information [8]. Events triggered by the very tight seeds are further processed for offline full event reconstruction. The AXOL1TL triggers performed stably during data-taking, as seen in Fig. 3.

The dataset triggered by AXOL1TL tends to be orthogonal to events triggered by the regular L1 Trigger menu, as seen in Fig. 4. This orthogonality highlights the novelty of the events collected by the L1 anomaly detection triggers. On closer examination of events triggered by both AXOL1TL and CICADA, we observe a preference for marking events with higher multiplicity as more anomalous. This reinforces the key role of these triggers in selecting events with multijet final states and BSM signatures, such as SUEPs, often overlooked by traditional rule-based algorithms.

The events collected by AXOL1TL are further examined for the viability of BSM physics searches using this data. As a preliminary check, the invariant mass distributions of pairs of jets, electrons, and photons are studied for potential biases or sculpting in the mass distributions. As observed in Fig. 5, these datasets seem to be free of any potential sculpting from trigger selections.

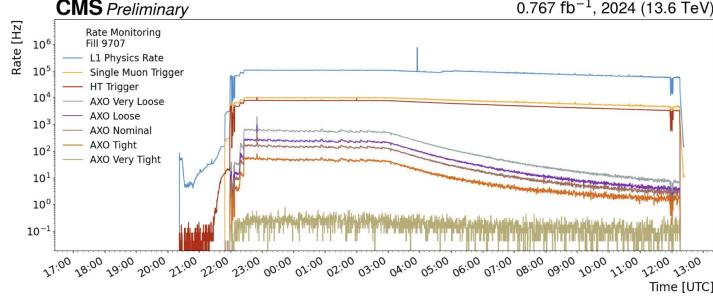


Figure 3: Global trigger rate monitoring time series over the course of data-taking in June 2024, showing the rates of AXOL1TL seeds.

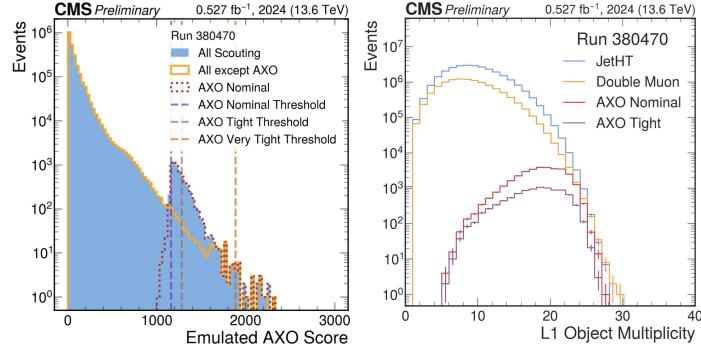


Figure 4: Left: Scores for all live AXOL1TL seeds and all events triggered by non-AXOL1TL HLT Scouting seeds, showing where the AXOL1TL contribution lies. Right: The distribution of AXOL1TL scores as a function of L1 object multiplicity.

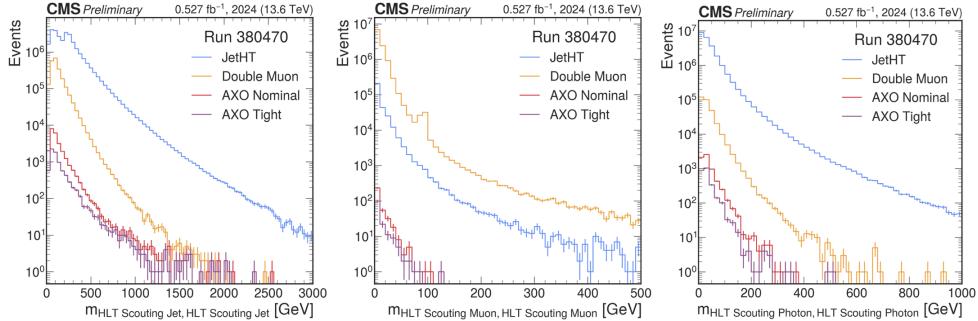


Figure 5: Invariant mass distributions of pairs of jets (left), muons (center), and photons (right) from objects reconstructed from data scouting as triggered by the AXOL1TL nominal trigger path.

4. Summary

This paper introduces two machine learning-based anomaly detection algorithms, AXOL1TL and CICADA, for real-time event selection in the CMS Level-1 (L1) Trigger at the Large Hadron Collider. These methods are designed to identify anomalous events without predefined selection criteria, improving sensitivity to potential new physics beyond the Standard Model. AXOL1TL uses a variational autoencoder to analyze L1 trigger objects, while CICADA employs a convolutional

autoencoder to process calorimeter energy deposits. Tested during CMS data-taking, AXOL1TL demonstrated stable performance and orthogonal event selection to traditional triggers, making it a promising tool for exploring novel physics signatures.

References

- [1] T. Arrestad et al. Fast convolutional neural networks on FPGAs with hls4ml. *Mach. Learn. Sci. Tech.*, 2(4):045015, 2021. doi: 10.1088/2632-2153/ac0ea1.
- [2] S. Chatrchyan et al. The CMS Experiment at the CERN LHC. *JINST*, 3:S08004, 2008. doi: 10.1088/1748-0221/3/08/S08004.
- [3] CMS collaboration. 2024 Data Collected with AXOL1TL Anomaly Detection at the CMS Level-1 Trigger. 2024. URL <https://cds.cern.ch/record/2904695>.
- [4] C. N. Coelho et al. Automatic heterogeneous quantization of deep neural networks for low-latency inference on the edge for particle detectors. *Nature Machine Intelligence*, 3(8): 675–686, June 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00356-5. URL <http://dx.doi.org/10.1038/s42256-021-00356-5>.
- [5] J. Duarte et al. Fast inference of deep neural networks in FPGAs for particle physics. *JINST*, 13(07):P07027, 2018. doi: 10.1088/1748-0221/13/07/P07027.
- [6] FastML Team. fastmachinelearning/hls4ml, 2023. URL <https://github.com/fastmachinelearning/hls4ml>.
- [7] E. Govorkova et al. Autoencoders on field-programmable gate arrays for real-time, unsupervised new physics detection at 40 mhz at the large hadron collider. *Nature Machine Intelligence*, 4(2):154–161, 2022.
- [8] A. Hayrapetyan et al. Enriching the Physics Program of the CMS Experiment via Data Scouting and Data Parking. 3 2024.
- [9] G. Hinton, O. Vinyals, and J. Dean. Distilling the Knowledge in a Neural Network. 3 2015.
- [10] V. Khachatryan et al. The CMS trigger system. *JINST*, 12:P01020, 2017. doi: 10.1088/1748-0221/12/01/P01020.
- [11] D. P. Kingma and M. Welling. Auto-encoding variational bayes, 2022. URL <https://arxiv.org/abs/1312.6114>.
- [12] Y. LeCun et al. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. doi: 10.1162/neco.1989.1.4.541.
- [13] R. Liu, A. Gandrakota, J. Ngadiuba, M. Spiropulu, and J.-R. Vlimant. Efficient and Robust Jet Tagging at the LHC with Knowledge Distillation. In *37th Conference on Neural Information Processing Systems*, 11 2023.
- [14] A. M. Sirunyan et al. Performance of the CMS Level-1 trigger in proton-proton collisions at $\sqrt{s} = 13$ TeV. *JINST*, 15:P10017, 2020. doi: 10.1088/1748-0221/15/10/P10017.