AutoDQM: A Machine Learning Approach to Data Quality Monitoring at CMS

Kaitlin Salyer

Goal: Develop a machine learning (ML) tool to highlight anomalies in detector output from the Compact Muon Solenoid (CMS) experiment at the Large Hadron Collider (LHC) to simplify the data quality monitoring (DQM) process.

Motivation: Previously, DQM shifters at CMS had to look at hundreds of plots of detector activity to determine if the hardware exhibited failures of any kind. This method was both time consuming and subject to human error: the shifters are not necessarily experts in the subsystems they are monitoring and they are looking for problems which are obscure and hard to spot by eye. To simplify this process, AutoDQM was designed to utilize statistical tests to compare histograms between data runs at CMS and flag outliers and highlight their anomalies. A visual example of the effectiveness of this approach is shown in Fig. 1. I seek to further develop this tool by implementing ML algorithms to make AutoDQM more flexible and adaptable to any subdetector system in CMS, or even in other experiments at the LHC.

Research Plan: I expect to work on this project half-time for the next 6 months (equivalent to 3 full-time months), with the other half of my time spent on physics analysis research. Our development team, lead by Dr. Indara Suarez (Boston University) and Chad Freer (Northeastern University), expects to use a variety of ML techniques to accommodate the various histogram types needed by CMS for DQM purposes. First, I will implement principle component analysis (PCA) for anomaly detection in 1D histograms for muon subsystems in CMS. I expect to complete the PCA module for these subsystems in the first 1-2 months of the fellowship. Subsequently, it will be necessary to develop a different ML module for anomaly detection in 2D histograms: possible avenues of interest to this end include autoencoders or generative adversarial networks. I expect, with the remainder of my 6 months, to first conduct studies to determine the best ML technique for general purpose use at CMS for anomaly detection in these 2D histograms, then to completely design a module using that method for implementation within the AutoDQM tool. I will be kept to this timeline by making regular progress updates not only to the AutoDQM development team, but also to the larger Muon ML research group at CMS. Furthermore, these ML developments will be designed using Python notebooks hosted through CERN's SWAN service, which will make them easily shareable with other CERN users. I will also be sure to document and make these codes available in a GitHub repository to be shared more broadly. It will be essential as well to design the ML modules to be easily retrainable for future updates to the detector and for applicability to other experiments beyond CMS.

Conclusions: Though AutoDQM was initially built as a tool for the muon subsystems at CMS, it is modular in structure, thus meaning it is easily adaptable to be used across all of CMS. We are already coordinating with the central DQM group to ensure this, and later I will collaborate with the PPD group to help AutoDQM become widely used in our experiment. Furthermore, this modularity means that this tool can be adapted outside CMS to other LHC experiments and beyond. This will ultimately streamline DQM activities and ensure that small anomalies in detector performance are not missed by shifters.

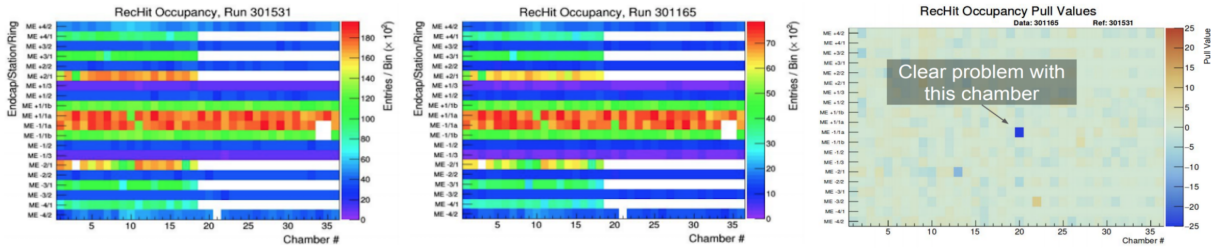

Figure 1: From left to right: the same histogram in a test run, a reference run, and what the current AutoDQM tool returns. As exemplified by these images, it is difficult to see much of a difference between the test and reference run, but AutoDQM's plot highlights the differences. It should be noted that this result from AutoDQM utilizes statistical tests instead of ML techniques.