Contribution ID: **62** Type: **Oral (16mins + 4 mins)**

Lean MLOps stack for development and deployment of Machine Learning models into an EPICS Control system

Wednesday, March 6, 2024 1:30 PM (20 minutes)

The ISIS Neutron and Muon Source is undergoing several upgrades to the control hardware, software, data acquisition and archiving systems. Machine learning systems are also being integrated into the control system. This not only requires the models to be high-quality but also to be maintained and kept up to date, especially in performance-critical applications. Each model incurs additional code-based maintenance, which can affect a project's longevity. For these reasons, we implemented a workflow for training and deploying models that utilises off-the-shelf, industry-standard tools such as MLflow, TF-Serve and Torch-Serve. We discuss the use of these tools; the adoption of lean paradigms and DevOps practices help optimise the developer throughput, maximise model quality and simplify monitoring and retraining. These tools and practices minimise the developer time spent on non-ML tasks, make it easy to track and compare model changes and performance as well as help improve the overall visibility of the projects across ML teams. We demonstrate how the models are automatically deployed and integrated into the EPICS control system and served to the end-user via Phoebus controls, which decreases the turn-around time for user feedback and together with the centralised model storage helps deliver and iterate the models rapidly. We discuss challenges and lessons learned along the development process, as well as a direction for future developments.

Primary Keyword

MLOps

Secondary Keyword

AI-based controls

Tertiary Keyword

Primary author: LEPUTA, Mateusz (UKRI-STFC-ISIS)

Presenter: LEPUTA, Mateusz (UKRI-STFC-ISIS)

Session Classification: Infrastructure / Deployment Workflows

Track Classification: Infrastructure / Deployment Workflows