Contribution ID: **39** Type: **Contributed Talk**

Robustness and interpretability of machine learning models applied to LHC data

Thursday 5 January 2023 18:15 (15 minutes)

Modern machine learning (ML) methods are widely used in LHC analyses, but considerably more time is invested in training ML models, than in understanding them. We present a small review of interpretation and explanation techniques relevant to ML classifiers used in collider experiments, and motivate why they should be consulted. Further, we present ongoing work on the related topic that is robustness, meaning how the output of a classifier is affected by changes in the input data. Different types of distribution shifts in data may affect the classifier output in nontrivial ways, which calls for systematic studies of model behaviour. We discuss plans for developing model-agnostic methods to quantise robustness under different types of perturbations in data.

Author: MAELAND, Steffen (Western Norway University of Applied Sciences)

Presenter: MAELAND, Steffen (Western Norway University of Applied Sciences)

Session Classification: Contributed Talks III

Track Classification: Dark matter experiments and experimental results