Contribution ID: **460** Type: **Sectional talk**

Data Shift Problem in Machine Learning for Particle Identification

Tuesday 8 July 2025 14:45 (15 minutes)

Particle identification (PID) is an essential step in the data analysis workflow of high-energy physics experiments. Machine learning approaches have become widely used in high-energy physics problems in general, and in PID in particular for the last ten years. Due to the fact that conventional algorithms of PID have poor performance in the high momentum range. However, due to the absence of ground-truth labels in experimental data, classifiers must be trained on Monte Carlo (MC) simulations. This creates a fundamental challenge: differences between the simulated and real data distributions known as data shift. It can significantly affect model generalization and performance. The impact of data shift was explored by comparing particle classification results across several MC datasets generated with different simulation settings. How the distributions of key features (momentum, energy, velocity, mass squared) vary between simulations was analyzed. The results highlight the need to carefully validate and adapt machine learning models to ensure reliable performance on data with potentially shifted distributions, especially in scenarios where real labels are unavailable.

Author: PAPOYAN, Vladimir (JINR & AANL)

Presenter: PAPOYAN, Vladimir (JINR & AANL)

Session Classification: Methods and Technologies for Experimental Data Processing