College of Integrative Sciences and Arts
Arizona State University

U.S. DEPARTMENT OF ENERGY

Evaluation of Machine Learning Methods on Simulated Tracks to Improve $K\,\pi$ identification in GlueX

Kevin Scheuer, Sebastian Cole, and Michael Dugger
College of Integrative Sciences and Arts, Arizona State University, Polytechnic

Introduction

The GlueX experiment began officially collecting physics data in 2016 with a ~12 GeV/c photon beam incident on a liquid hydrogen target. GlueX data from 2016 to 2019 is plagued by a difficulty to identify kaons, ending with the integration of the DIRC (Detection of Internally Reflected Cherenkov light) detector into the beamline in 2020. Machine learning algorithms from Python's sklearn library have been applied to a Monte Carlo produced data set of 500,000 K^+ and 500,000 π^+, in order to improve the identification of kaons beyond 1.0 GeV/c in momentum. Initial tests of the models on the GlueX data have also begun.

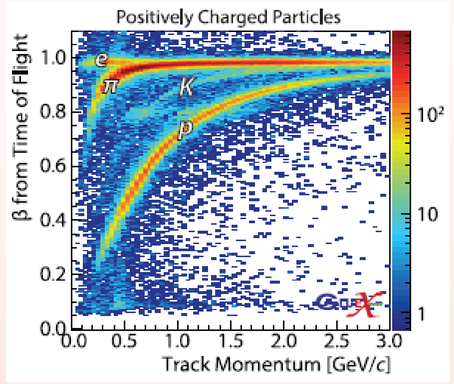

Figure 1: Beta vs momentum plot [1].

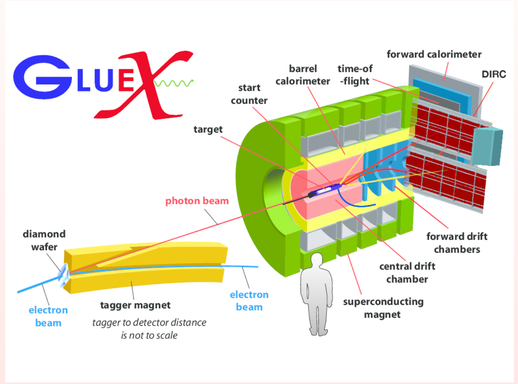

Figure 2: Schematic of GlueX detector [1].

Feature Selection

Pearson correlation coefficients were calculated to measure the linear correlation between selected features of the Geant3 produced charged tracks. These features come from detector response and reconstruction of charged tracks in the detector. They are used to train the four algorithms for analysis.

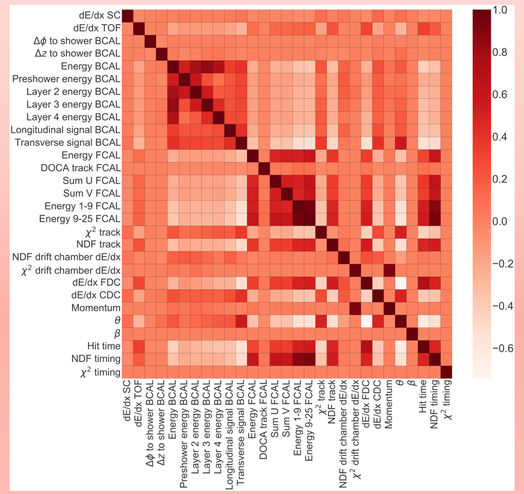

Figure 3: Heatmap of the Pearson correlation coefficients for selected features.

Cross Validation

Using the selected features, four classifiers from Python's sklearn library were applied to best determine what machine learning algorithm may provide greater $K\,\pi$ discrimination between tracks. The algorithms are:

- Gradient Boosted Decision Tree (GBDT),
- Random Forest (RF)
- Multi-Layer Perceptron (MLP)
- Adaptive Boosted Decision Tree (ADBT).

K-fold cross validation with 5 folds was applied to test the generalizability of the analysis models to application with the GlueX data.

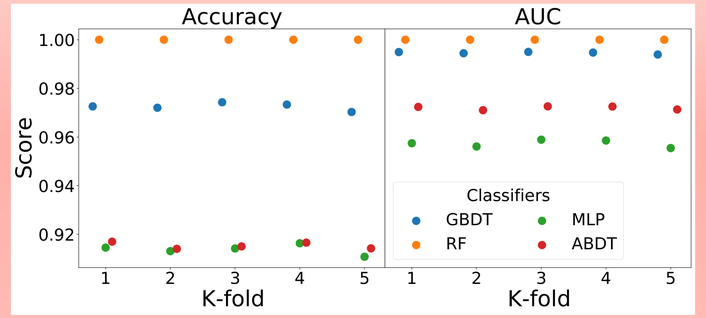

Figure 4: Results of 5-fold cross validation for each fold.

Evaluation

The mean and standard deviation of the accuracies and the area under the receiver operating characteristic curve (ROC AUC) were produced to provide a measure of a model's performance and possible error. The GBDT performed the best with an accuracy of 97.2% and AUC of 0.995. The left plot in Figure 6 shows the RF performing better, but it produced bias in the application to real data. The results of application of the GBDT model to low intensity 2017 GlueX data are promising, but further work is needed in tuning the parameters of the model.

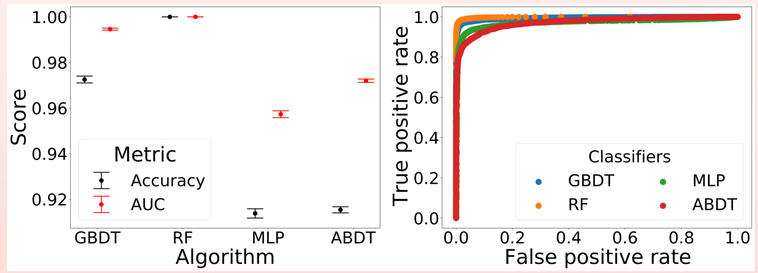

Figure 5: Mean accuracy and AUC scores, with ROC curve results of each classifier.

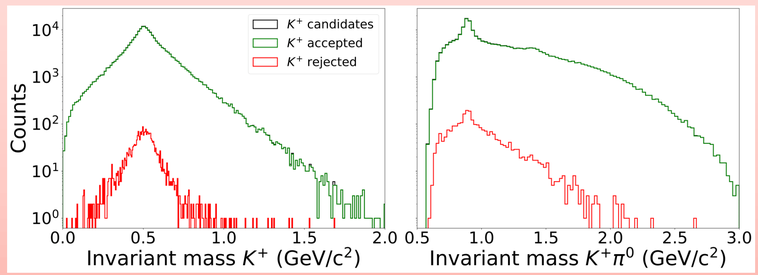

Figure 6: Invariant mass histograms of $K^+$and $K^+\pi^0$ for $\gamma\, p \rightarrow p\, K^+K^-\pi^0$ events in GlueX 2017 low intensity data before and after GBDT particle identification.

Future directions

With the success of the GBDT, effort is being made to improve the model with training on a larger data set, hyperparameter tuning, and application to more GlueX data.

[1] Ghoul H. *et al.* (2016) 'First Results from The GlueX Experiment', arXiv. Available from: https://arxiv.org/abs/1512.03699