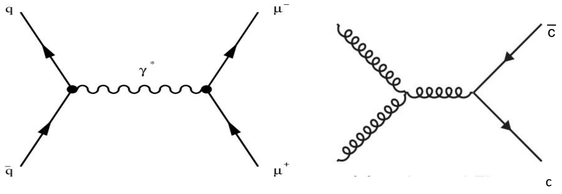
Using Machine Learning to Improve Dilepton Signal Extraction

Gabriel Rodriguez, Bishoy Dongwi, Charles Naïm
Department of Physics and Astronomy

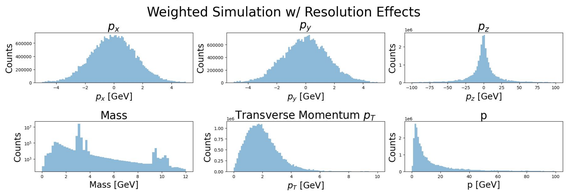
Introduction

Dilepton production within p+p collisions are the result of several different processes: the Drell Yan process, J/ψ, Υ, open charm, open bottom, and light meson decays. Isolating individual processes to study is difficult due to signal overlap. In response to this problem, machine learning techniques can be used to improve signal-to-background data extraction. By training a machine learning model to categorize dilepton pairs, researchers can easily isolate individual processes despite the signal overlap. These new techniques will provide researchers with more accurate datasets, helping them to study these processes further.

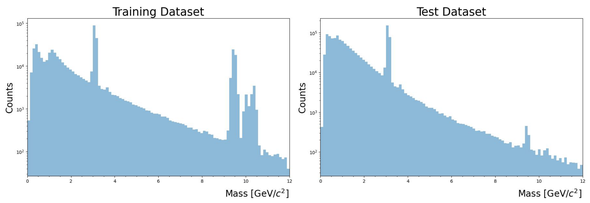


Simulation

Before a machine learning model can make predictions, it must first be trained on a dataset. The training datasets were generated using PYTHIA 8.315 to simulate p+p collisions at a center-of-mass energy of √s = 200 GeV.

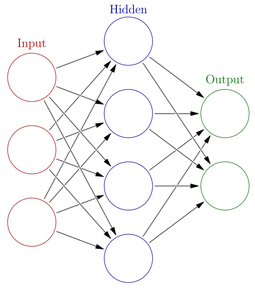


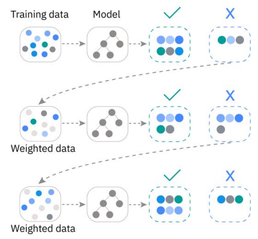
Two datasets were generated using PYTHIA: one for training and one for testing. The training dataset was made by individually simulating each process, normalizing their contributions, and then combining the results to create a balanced dataset in which all processes are equally represented. In contrast, the test dataset was generated by simulating all processes simultaneously, creating a more realistic result that offers a more accurate assessment of the model's performance.



Methodology

Initial attempts to use machine learning for classification involved a neural network; however, the neural network was inefficient due to its long runtime and poor accuracy. The approach was then changed from using a neural network to using gradient boosting with the XGBoost Python library. Gradient boosting is designed and optimized for classification, using iteration and several weak decision trees to make predictions. Shifting to gradient boosting increased accuracy from 20% to 90% and reduced runtime significantly.

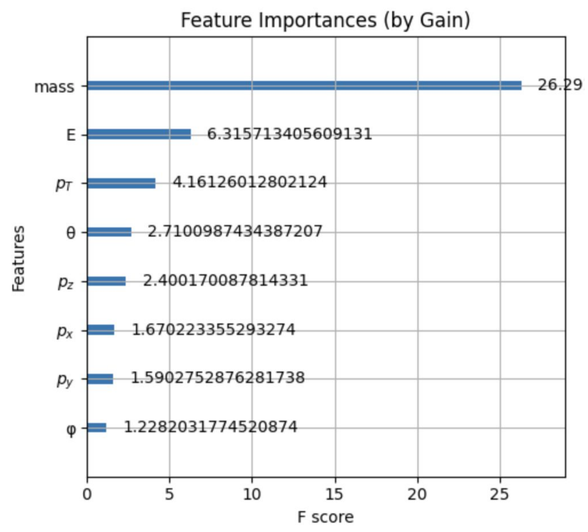


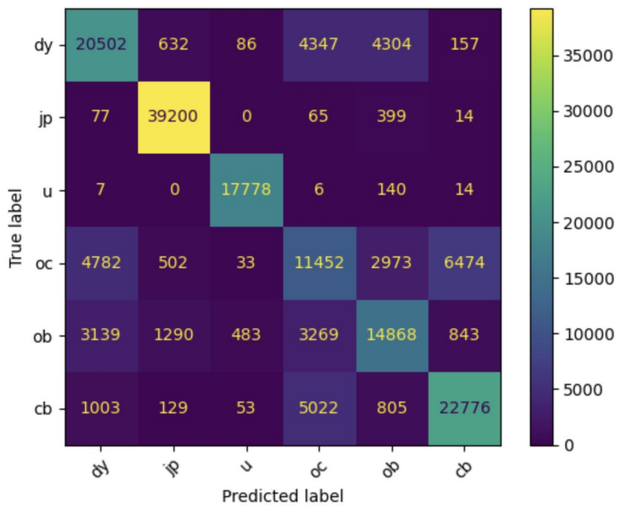


Initial results showed that the model concentrated on the most frequent processes only, leading to lower accuracy for less represented ones. To address this, the training dataset was constructed with equal proportions of dilepton pairs from each process, aiming to promote balanced predictive performance. The test dataset, by contrast, was designed to reflect a realistic distribution of events, providing a more accurate assessment of the model's effectiveness under practical conditions.

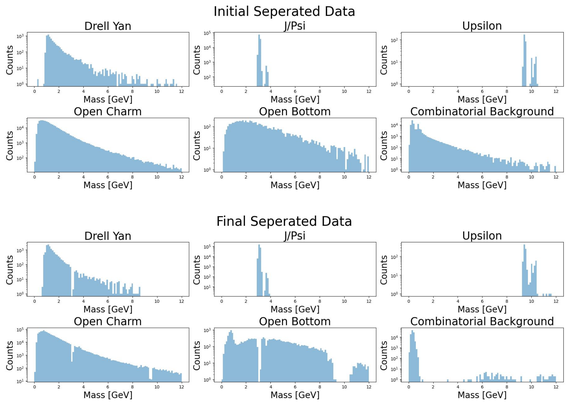
	precision	recall	f1-score	support
Drell Yan	0.69	0.68	0.69	30028
J/ψ	0.94	0.99	0.96	39755
Υ	0.96	0.99	0.98	17945
Open Charm	0.47	0.44	0.45	26216
Open Bottom	0.63	0.62	0.63	23892
CB	0.75	0.76	0.76	29788
accuracy			0.76	167624
macro avg	0.74	0.75	0.74	167624
weighted avg	0.75	0.76	0.75	167624

Model Training Results





Model Prediction Results



The trained model is able to predict the process responsible for creating each dilepton pair, and classify the pairs by process for further analysis. Pictured above are the dilepton pairs plotted by process, first separated manually for comparison, and below separated by the model using its predictions.

Conclusion

Preliminary findings demonstrate that machine learning techniques can be effectively applied to signal processing, particularly for distinguishing signal data from background noise. The model developed and employed here effectively isolates signals across different processes, demonstrating reliable performance in signal extraction.

Test Accuracy: 0.8266

	precision	recall	f1-score	support
Drell Yan	0.31	0.31	0.31	13583
J/ψ	0.94	0.98	0.96	225621
Υ	0.60	0.94	0.74	586
Open Charm	0.84	0.91	0.88	800535
Open Bottom	0.19	0.23	0.21	12013
CB	0.58	0.33	0.42	180031
accuracy			0.83	1232369
macro avg	0.58	0.62	0.59	1232369
weighted avg	0.81	0.83	0.81	1232369