CPP

1. Data Summary

The nChargedTracks value counts only pions (PID=8 or 9) where the track projection contacts the first FMWPC layer. The nFMWPCMatchedTracks values do not exclude tracks whose projections do not meet the first FMWPC layer. As a result, the number of matched tracks may exceed "nChargedTracks." The FMWPC closest wire1 and FMWPC_dist_closest_wire1 variables include information on the track striking the first layer, at least partially.
The fact that the same wire is listed as the closest hit for numerous tracks is due to just one wire being hit in the layer. What should be looked at is the value in FMWPC_dist_closest_wire. If the number is big, it indicates that the track is not well matched to it.

1.1. nChargedTracks:

The number of reconstructed tracks that met the following conditions:

1. Had either a pi+ or pi- mass hypothesis
2. Track projected cleanly to the FCAL
3. Track projected cleanly to the first layer of the FMWPC

1.2. nFCALShowers:

Number of reconstructed FCAL showers. Showers, by default, require a minimum of two blocks. Single blocks where no neighbors are hit are, therefore, not counted. There is also a 350MeV minimum energy deposit. This means that there will be no equivalent "shower" for minimal ionizing particles (MIPs) that travel directly through a single block, depositing a little amount of energy. This is true for the vast majority of MUons and certain PIons.

1.3. nFCALHits:

The total number of FCAL hits. This is the total number of blocks struck throughout the event. Technically, two hits might be included in a single block if they were far enough apart in time, but this is highly unlikely. Because there is no time limit on these, there may be accidentals from unrelated events that occur near to the triggering event. This is absent in the simulated data but will be present in the real data.

1.4. nMWPCHits:

number of FMWPC wire hits in the event.

1.5. nMWPCMatchedTracks:

Number of "matched" tracks in the event. This is the number of reconstructed tracks that were one of a e+,e-,pi+,pi- mass hypothesis. Other criteria were not used. As a result, tracks that were not counted in the nChargedTracks field above will be tallied here.

All values are variable-sized arrays containing nFMWPCMatchedTracks items. In the Hall-D code, they correspond to the DFMWPCMatchedTrack reconstructed data items. The term

"matched" refers to including information from the track as well as hits/clusters in the FCAL and FMWPC detectors.

1.6. FMWPC_pid:

The mass hypothesis for each matched track. These should be one of:

2 = positron
3 = electron
8 = pi+
9 = pi-

1.7. FCAL_E_center:

The center position where the calibrated energy is projected to hit the FCAL.

While this is a calibrated hit, it excludes things like depth corrections applied to blocks that are part of a cluster. Furthermore, because the pion and muon particles do not create EM showers, there is a lot of energy deposited in the downstream end of the block that is less attenuated than typical shower energy. As a result, their "apparent" energy will be higher.

Finally, if the track projection points to a different block than the one that was actually hit, this can easily be 0. So, this is likely to be of limited utility.

1.8. FCAL_E_3x3:

The sum of the 3x3 group of FCAL blocks centered around the FCAL E center block.

This is just meant to look at a small portion of the FCAL along the projected track. It is straight sum of the FCAL hits with no depth correlation etc.

1.9. FCAL_E_5x5:

The sum of the 5x5 group of FCAL blocks centered around the FCAL E center block.

Note – it is likely that the ratio of 3x3 or/and 5x5 will be a good fit to train on since it indicates how contained the shower is.

1.10. FMWPC_Closest_wire:

This is the closest hit wire in the FMWPC to the projected track for the FMWPC layer.
Some layers will have no wires hit in the FMWPC's. As a result, a value of -1000 will be returned, indicating that the appropriate FMWPC Closet wire was not struck.

1.11. FMWPC_dist_closest_wire:

The distance between the layer's track projection and the closest hit wire. This is in wire units, and the wire spacing is little over 1cm. This defaults to 1.0E6 if no wires were hit in the layer.

1.12. FMWPC_Nhits_cluster :

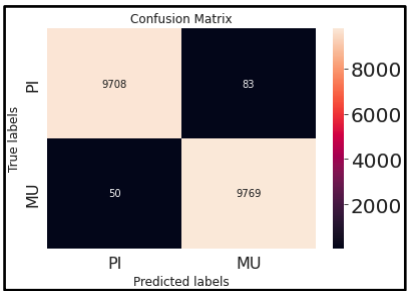
The number of contiguous wires hit that include FMWPC closest wire. This would, in theory, be a measurement of the size of a hardonic shower passing over the FMWPC layer.

2. Model evaluation

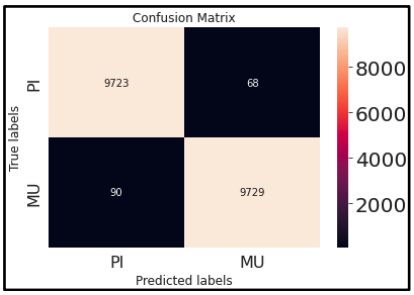
2.1. Confusion matrix:

Various metrics are calculated to evaluate the classifier. The accuracy measure is the ratio of correctly identified events to the total number of occurrences in the data. Metrics generated from the Confusion Matrix are explored in order to better understand the classifier's performance.

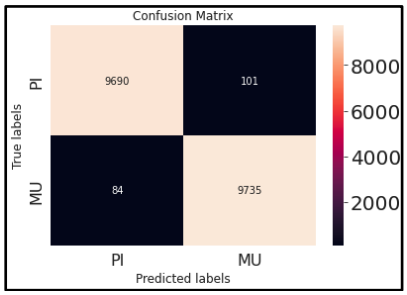
The True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) values are used to evaluate the Confusion Matrix (FN). True positive (TP) refers to the amount of accurately predicted PION occurrences in the image above (figure 1). FP refers to the number of incorrectly anticipated pion events is referred to as muon events. FN stands for the number of muon events that were incorrectly predicted as PION events, whereas TN stands for the number of muon events that were correctly anticipated.



XGBoost algorithm 1.3.0



Random Forest algorithm



MLP classifier

2.2. Classification report:

Data scientists frequently consult classification reports to assess all of the options for improving our model and presenting the best model report. It's used to demonstrate the trained classification model's accuracy, recall, F1 score, and support.

$$
\begin{aligned}
precision &= \frac{TP}{TP+FP} \\
recall &= \frac{TP}{TP+FN} \\
F1 &= \frac{2 \times precision \times recall}{precision + recall} \\
accuracy &= \frac{TP+TN}{TP+FN+TN+FP} \\
specificity &= \frac{TN}{TN+FP}
\end{aligned}
$$

classification_ report

	precision	recall	f1-score	support
0	0.994	0.996	0.995	17526
1	0.996	0.994	0.995	17820
accuracy			0.995	35346
macro avg	0.995	0.995	0.995	35346
weighted avg	0.995	0.995	0.995	35346

XGBoost algorithm 1.3.0

classification_ report

	precision	recall	f1-score	support
0	0.993	0.991	0.992	9813
1	0.991	0.993	0.992	9797
accuracy			0.992	19610
macro avg	0.992	0.992	0.992	19610
weighted avg	0.992	0.992	0.992	19610

Random forest algorithm

classification_ report

	precision	recall	f1-score	support
False	0.988	0.992	0.990	9781
True	0.992	0.988	0.990	9829
accuracy			0.990	19610
macro avg	0.990	0.990	0.990	19610
weighted avg	0.990	0.990	0.990	19610

MLP classifier

2.3. ROC Curve:

A receiver operating characteristic curve (ROC curve) is a graph that shows how well a classification model performs across all categorization levels. Two parameters are shown on this curve:

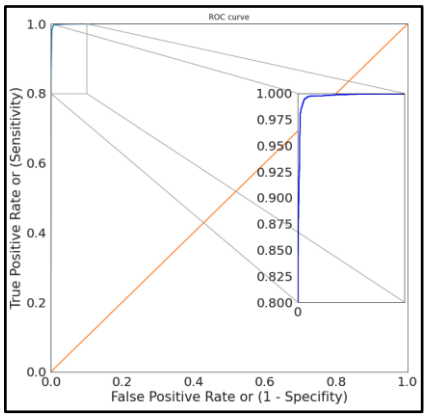
- True Positives Rate [a synonym for Recall]

$$TPR = \frac{TP}{TP + FN}$$

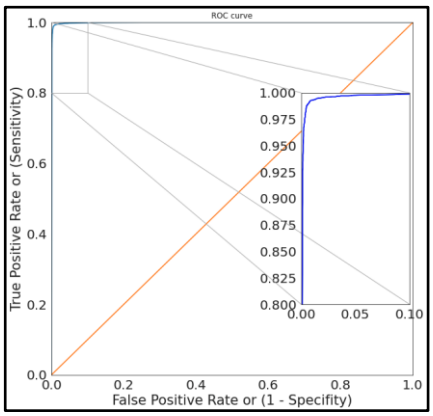
- False Positives Rate

$$FPR = \frac{FP}{FP + TN}$$

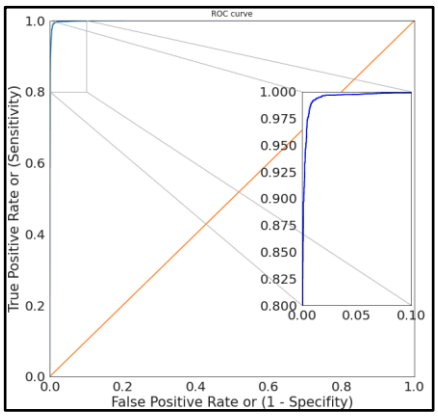
TPR vs. FPR at various categorization criteria is plotted on a ROC curve. As the classification threshold is lowered, more items are classified as positive, resulting in an increase in both False Positives and True Positives.



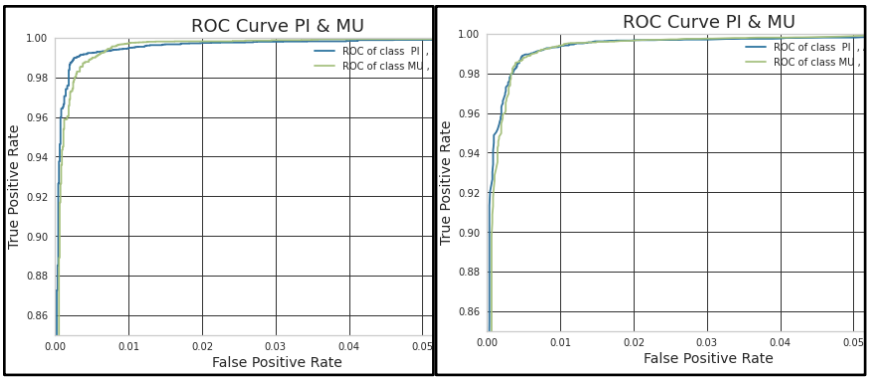
XGBoost algorithm



Random forest algorithm



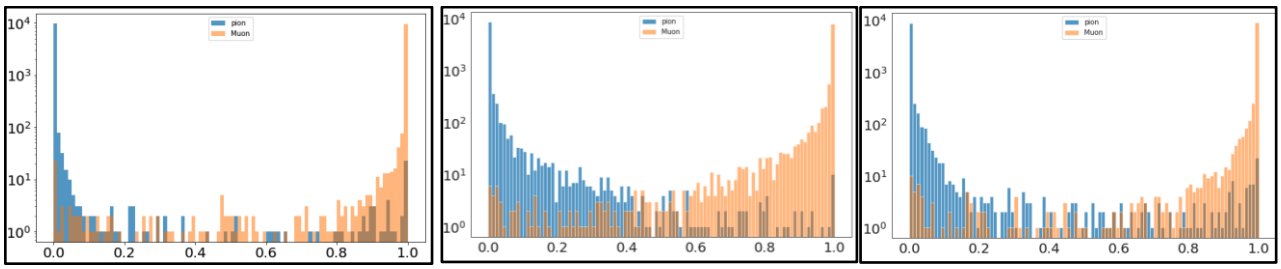
MLP classifier



2.4. Response of Machine Learning models:

The classifier's response is the likelihood that an event is a signal event, in this case a Pion event. For a perfect classifier, all signal events should have a reaction of 1 and all background events should have a response of 0, resulting in an easy separation.

However, because the classifiers under consideration here are imperfect, a cut must be made, establishing the conditions for a signal or background event. This is referred to as the threshold value. For all of the classifiers tested, a threshold of 0.5 is used, which means that any event with a reaction greater than or equal to 0.5 is classed as a signal event, and any event with a response less than or equal to 0.5 is categorized as a background event.



XGBoost algorithm 1.3.0 *Random forest algorithm* *MLP classifier*

3. Links of Most important notebooks -

1. Model's notebook. (Matched_root_data_with_background) (All Imp. Models at one place with HPO)-

https://github.com/JeffersonLab/jlab_datascience_pid/blob/main/cpp/notebooks/new%20data%20notebooks/new%20data_Matched_w_Background_apr_16th_pion%20and%20muon_XGB_RF_auto_encoders_conv1D_MLP---%20FINAL%20----.ipynb

Pickle file after data prep (used in all the matched_data notebook)-

new_balanced_df = pd.read_pickle ("/work/data_science/kalra/pickle_file/final_df_with_calculated_features_matched_background.pkl")

2. SNGP model notebook -
https://github.com/JeffersonLab/jlab_datascience_pid/blob/main/cpp/notebooks/new%20data%20notebooks/SNGP-diff_features_matched_tracks.ipynb

3. Data Prep Matched .py file -
https://github.com/JeffersonLab/jlab_datascience_pid/blob/main/cpp/data_prep/matched_root_dataprep.py

4. Data Prep Matched Notebook -
https://github.com/JeffersonLab/jlab_datascience_pid/blob/main/cpp/notebooks/data_prep_notebook/matched_root_dataprep.ipynb

5. Data analysis of Matched data with Electro-Magnetic Background -
https://github.com/JeffersonLab/jlab_datascience_pid/blob/main/cpp/notebooks/new%20data%20notebooks/new%20data_2nd%20root%20file_apr_16th_pion%20and%20muon-200000-data%20analysis_w_Background.ipynb

6. Final best models used - (tflite_model, xgb_root_model, MLP_h5_model)
https://github.com/JeffersonLab/jlab_datascience_pid/tree/main/cpp/model/2022-05-03%20final_test_models

7. Data analysis of Matched data without Background -
https://github.com/JeffersonLab/jlab_datascience_pid/blob/main/cpp/notebooks/new%20data%20notebooks/new%20data_2nd%20root%20file_feb_2nd_pion%20and%20muon-200000-data%20analysis-no_Background.ipynb

8. Slides by andrew
https://halldweb1.jlab.org/wiki/images/4/4f/DRAFT_CPPNPP_RunPlanPrep_GlueXCollab_May2022.pdf

9. Slide deck on step by step procedure -
https://teams.microsoft.com/_#/files/CHARGED%20PION%20POLARIZABILITY%20(CPP)%E2%80%8B%20Study?threadId=19%3A77c11d4ed18443a68983112d3bad059b%40thread.tacv2&ctx=channel&context=presentation&rootfolder=%252Fsites%252FDataScience-ChargeParticlePolarityStudy%252FShared%2520Documents%252FCharge%2520Particle%2520Polarity%2520Study%252Fpresentation

10. Additional link that will help us -
https://surface.syr.edu/cgi/viewcontent.cgi?article=1846&context=etd
https://www.newscientist.com/article/mg12717284-700-muons-pions-and-other-strange-particles/
https://www.worldscientific.com/doi/abs/10.1142/S0217732308023797
https://indico.cern.ch/event/51276/contributions/2034213/attachments/967066/1373360/305_Nappi.pdf
http://bes.ihep.ac.cn/bes3/phy_book/book/phy/ParticleID.pdf
https://halldweb.jlab.org/DocDB/0049/004905/001/Overview_of_experiment_and_setup.pdf
https://halldweb.jlab.org/DocDB/0046/004670/002/CPP%20jeopardy%20presentation.pdf

4. Old data

https://github.com/JeffersonLab/jlab_datascience_pid/blob/main/cpp/notebooks/new%20data%20notebooks/new%20data_jan%2019th_pion%20and%20muon.ipynb

Pickle file -

```
final_df = pd.read_pickle('/work/data_science/kalra/pickle_file/final_df_19th_jan_10k.pkl')
df_pi = pd.read_pickle('/work/data_science/kalra/pickle_file/df_pion_19th_jan_10k.pkl')
df_mu = pd.read_pickle('/work/data_science/kalra/pickle_file/df_muon_19th_jan_10k.pkl')
```