RICH AI

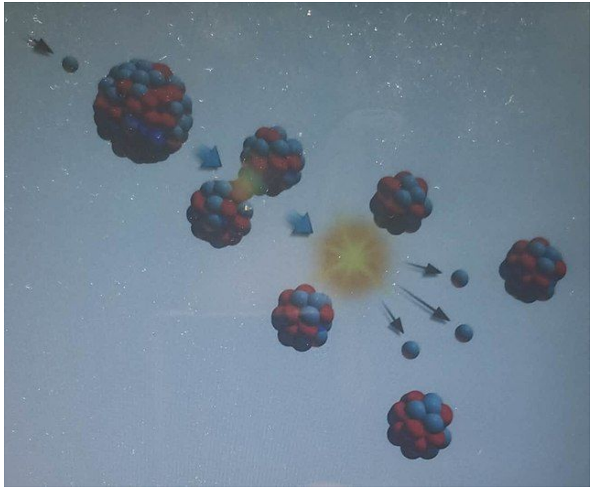
Mukund Iyer, Nico Van den Hooff,
Rakesh Pandey, Shiva Jena

Agenda

1) Background and scope
2) Data
3) Machine learning approach
4) Evaluation metrics
5) Timelines

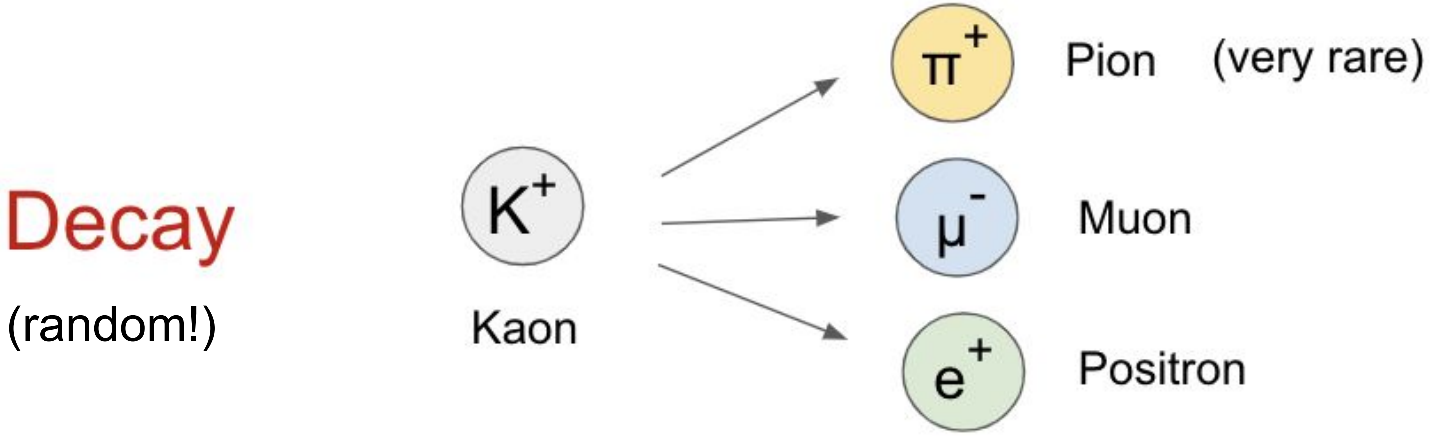
1. Background and scope

The NA62 particle physics experiment at CERN detects sub-atomic particles



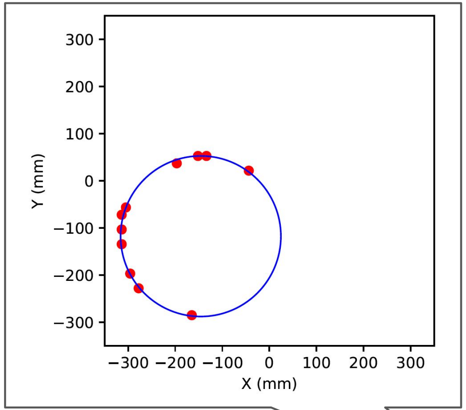
Geneva, Switzerland

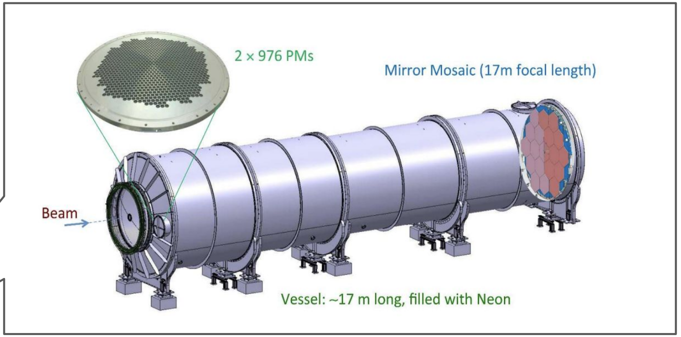
The challenge is to match the experimental results to the standard model



There are 3 main particles generated in the NA62 experiment

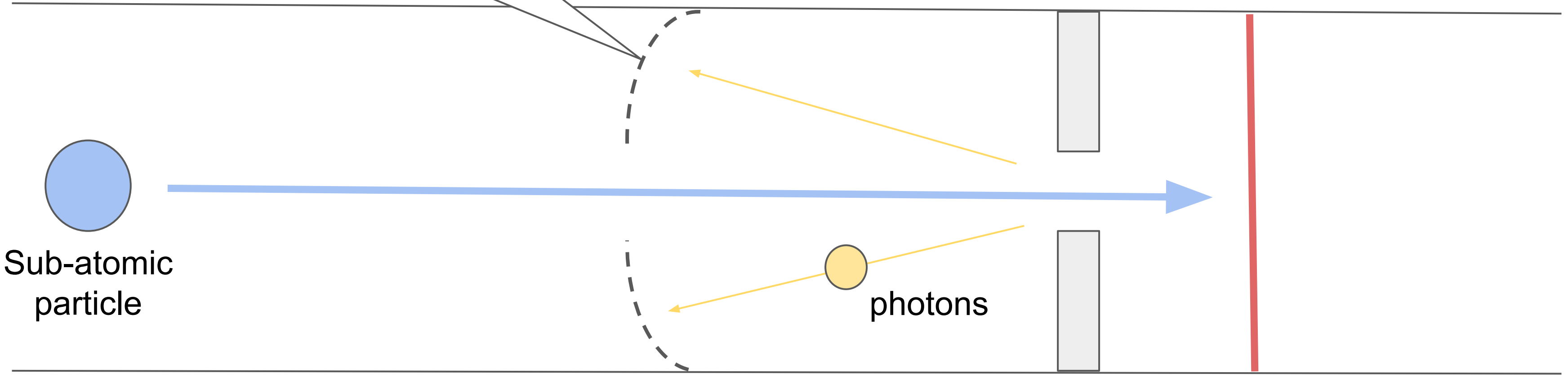
The RICH detector is used to produce light




2 × 976 PMs
Mirror Mosaic (17m focal length)
Beam
Vessel: ~17 m long, filled with Neon

RICH detector

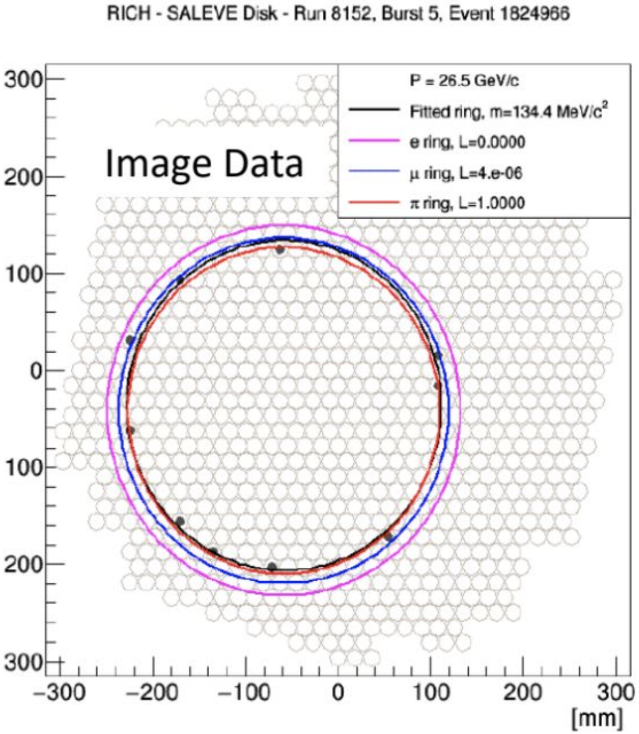
Hodoscope



Sub-atomic particle

photons

Each particle produces a different size ring.. image classification!



Current method

- MLE used to fit rings analytically

Challenges

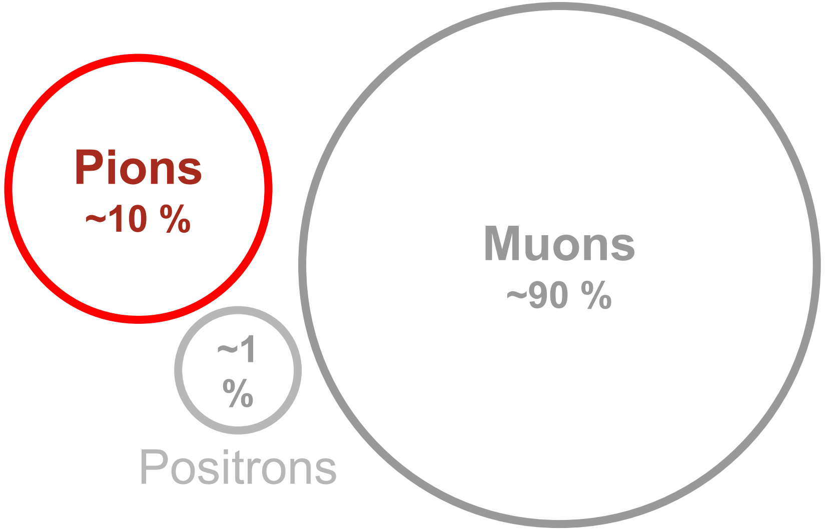
- Pion decay is very rare
- Capture as many pion decays as possible
- Reduce the number of incorrect classifications

2. Data

Data: *description*

Source	NA62 datasets from RICH detector
Type	Labelled datasets
Format	HDF5
Volume	~11.5 million examples*

Class Imbalance*



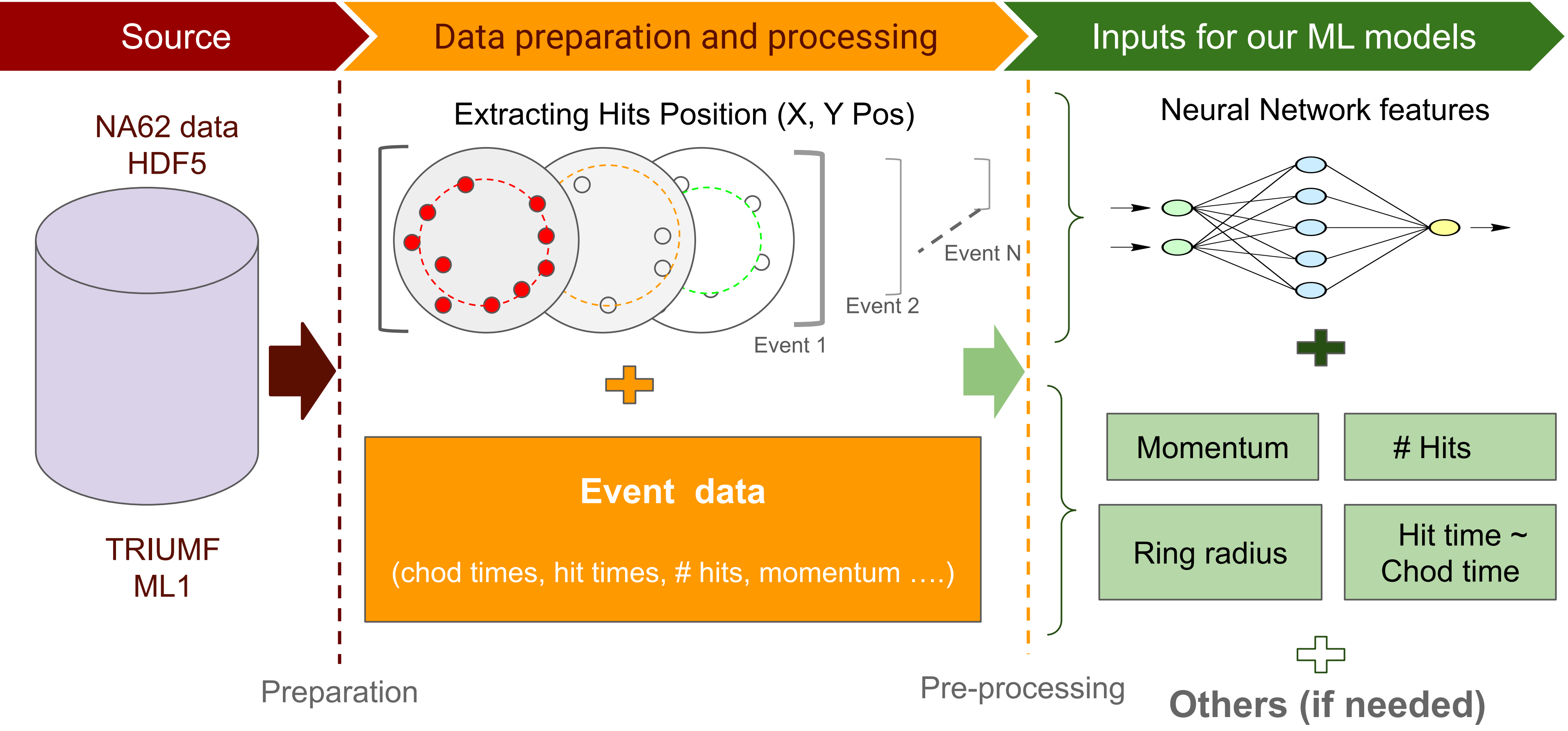
Pions
~10 %

Muons
~90 %

~1 %
Positrons

*Source: As per the initial datasets shared by TRIUMF for 2018 (B & E); Other 4 years data would be shared on need basis
https://github.com/TRIUMF-Capstone2022/RICHPID/tree/main/docs

Data: *flow*



Source

Data preparation and processing

Inputs for our ML models

NA62 data
HDF5

TRIUMF
ML1

Extracting Hits Position (X, Y Pos)

Event data

(chod times, hit times, # hits, momentum ….)

Neural Network features

Momentum

Hits

Ring radius

Hit time ~
Chod time

Others (if needed)

Preparation

Pre-processing

3. Machine Learning Approach

Machine learning approach

Task	Build a **classification** model: pions, muons, positrons
Goals	*Goal 1*: **Maintain** pion efficiency of **95%** (TP rate) *Goal 2*: **Reduce** pion/muon misclassification by **10x**
Baseline	Gradient Boosted Tree (XGBoost or LightGBM)
Deep Learning	*Model 1*: PointNet or PointNet++ *Model 2*: Graph CNNs
Deep Learning Implementation	PyTorch Geometric

4. Evaluation metrics

Evaluation: ROC curve to compare classifiers

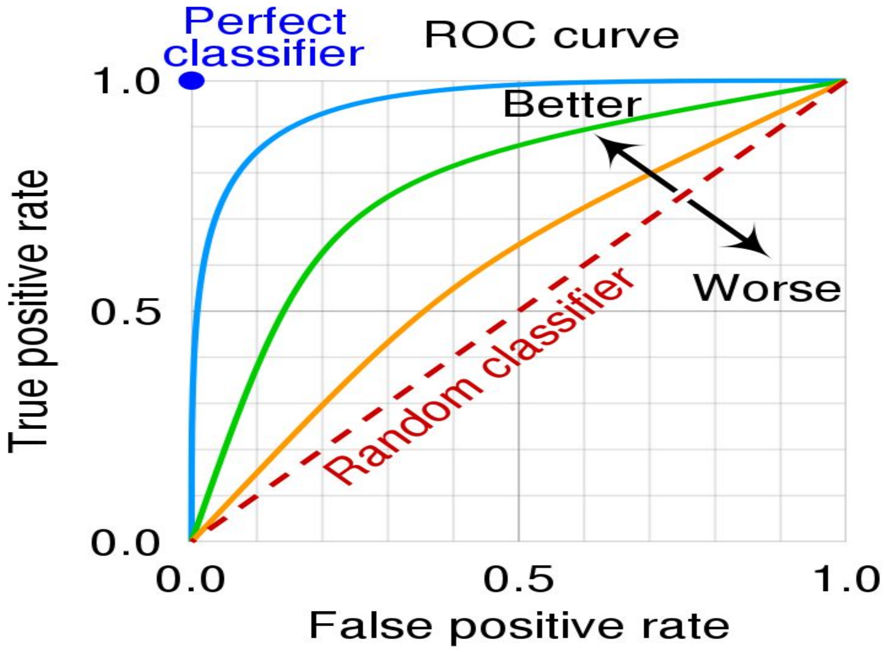


**Image source: wikipedia

ROC Curve

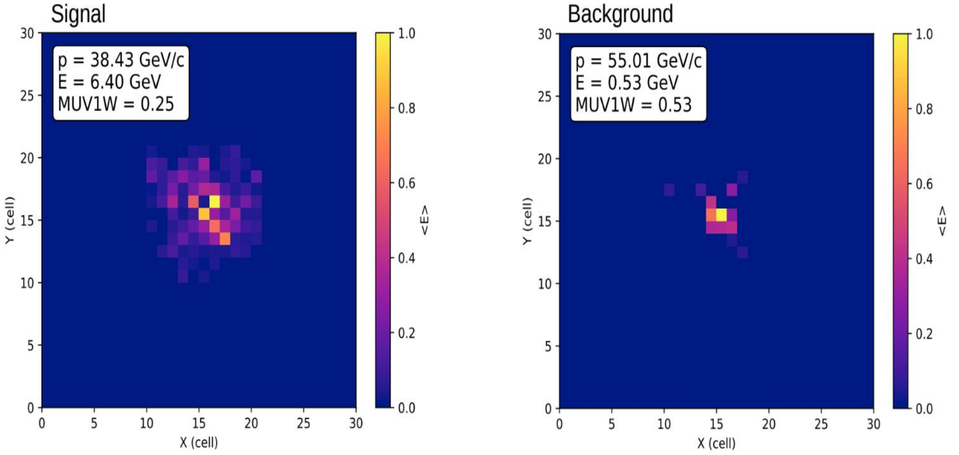
- Useful in comparing different classifiers
- Diagonal line represents random classifier
- Models below diagonal lines are worse than random classifier
- Area under ROC curve (AOC) - ability of a classifier to distinguish between classes

Evaluation: Pion efficiency

Fraction of matched detected pions out of the all the pion particles.

	Predicted PION	Predicted NOT PION
IS a PION	True Positives	False Negatives
NOT a PION	False Positives	True Negatives

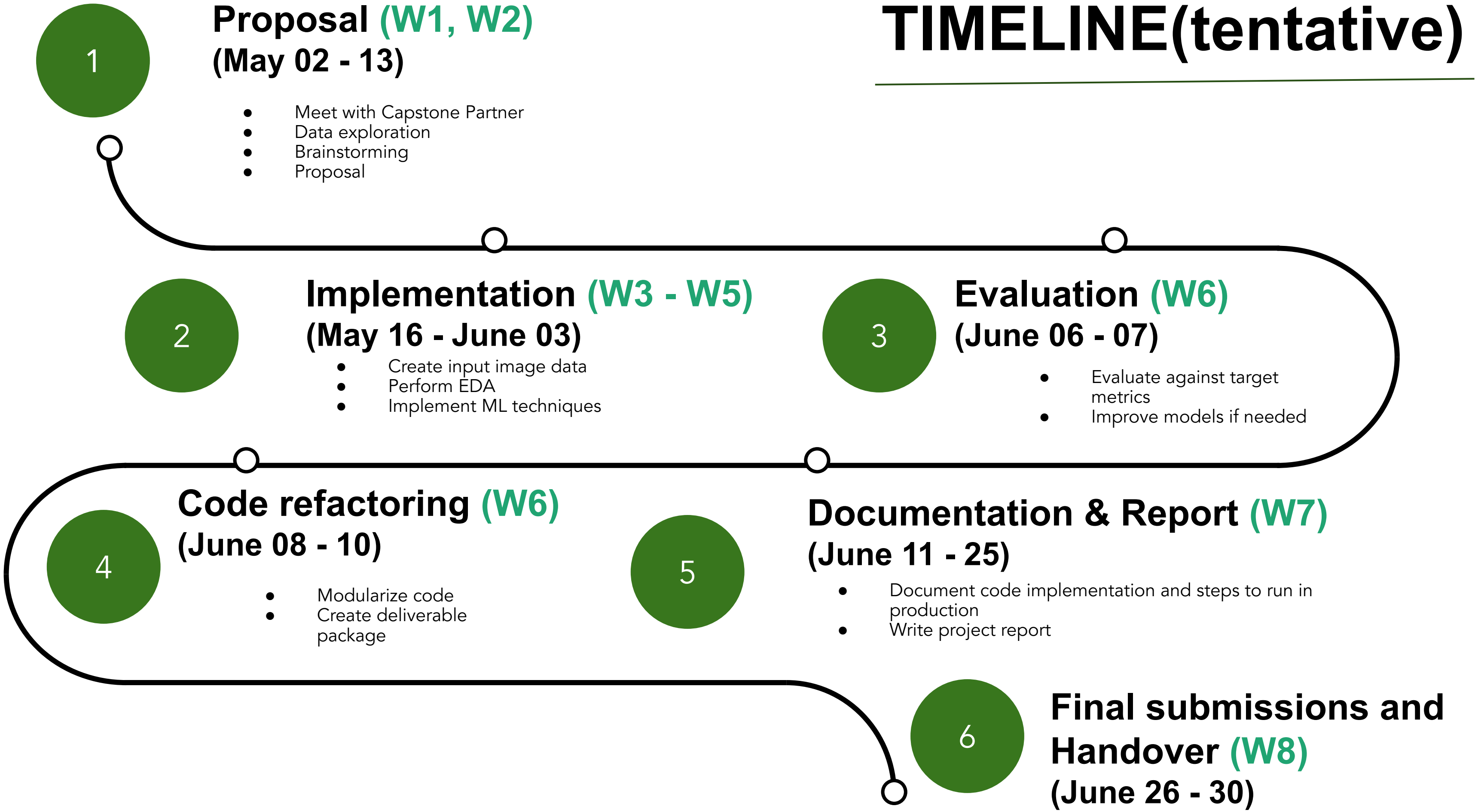
$$\text{Pion efficiency} = \frac{\text{\#pions detected}}{\text{\#pions}}$$



- Need 95% efficiency for the signal

5. Timelines

TIMELINE(tentative)



1

Proposal (W1, W2)
(May 02 - 13)

- Meet with Capstone Partner
- Data exploration
- Brainstorming
- Proposal

2

Implementation (W3 - W5)
(May 16 - June 03)

- Create input image data
- Perform EDA
- Implement ML techniques

3

Evaluation (W6)
(June 06 - 07)

- Evaluate against target metrics
- Improve models if needed

4

Code refactoring (W6)
(June 08 - 10)

- Modularize code
- Create deliverable package

5

Documentation & Report (W7)
(June 11 - 25)

- Document code implementation and steps to run in production
- Write project report

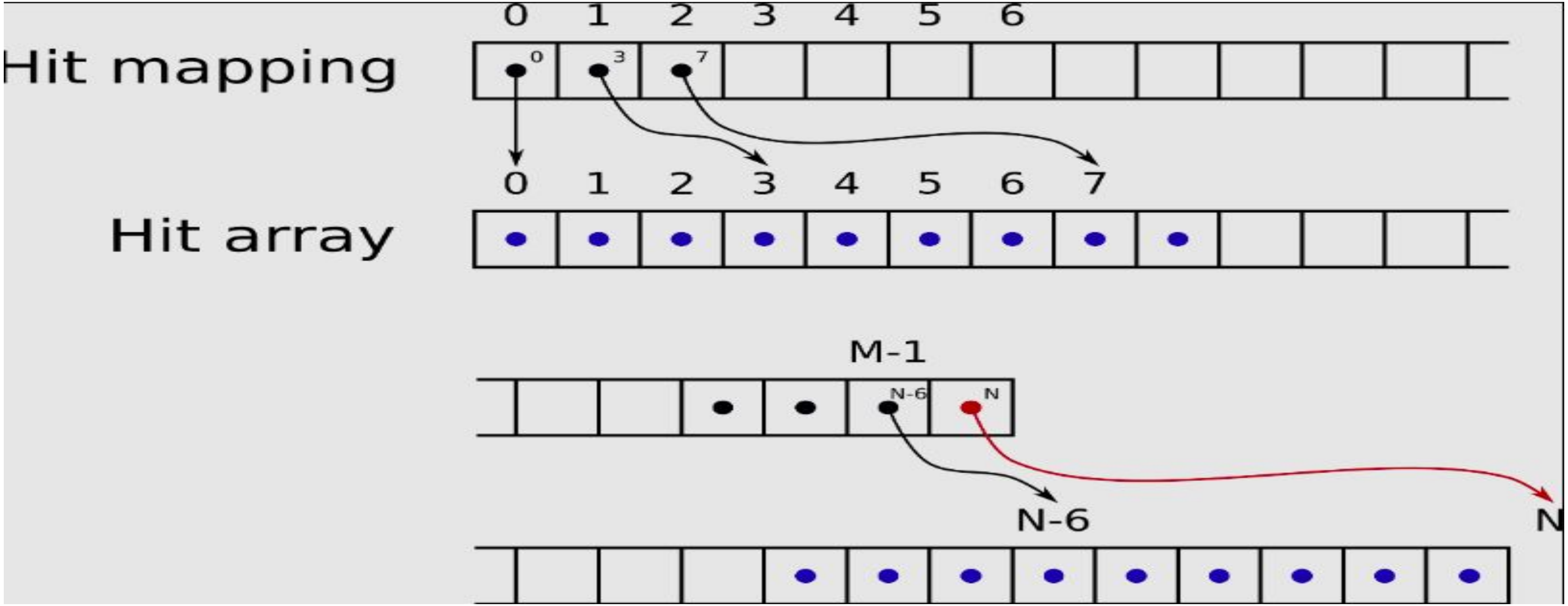
6

Final submissions and Handover (W8)
(June 26 - 30)

THANK YOU

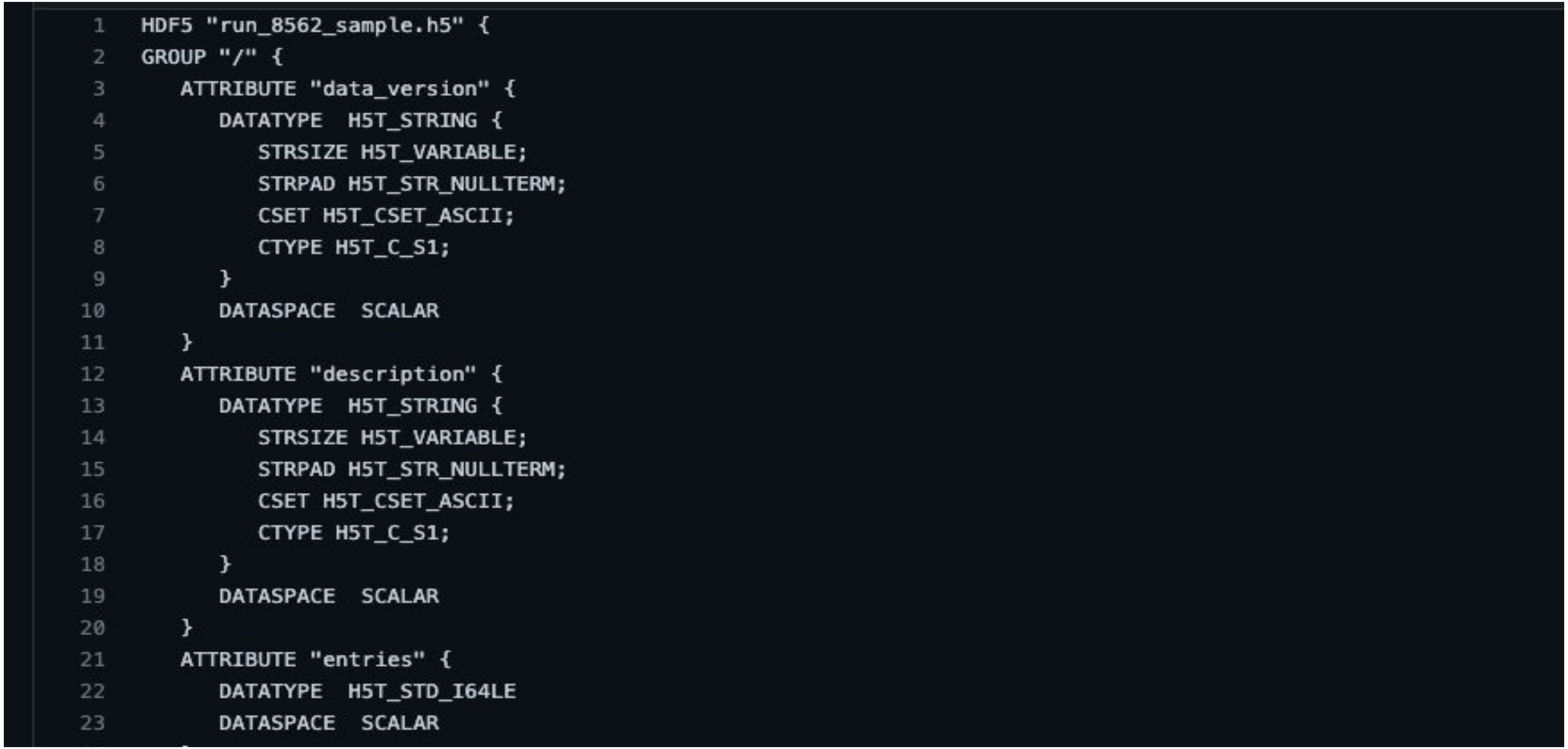
APPENDIX

Data - structure and *indexing*



*Source: https://github.com/TRIUMF-Capstone2022/RICHPID/tree/main/docs

HDF5 format



```
HDF5 "run_8562_sample.h5" {
GROUP "/" {
   ATTRIBUTE "data_version" {
      DATATYPE  H5T_STRING {
         STRSIZE H5T_VARIABLE;
         STRPAD H5T_STR_NULLTERM;
         CSET H5T_CSET_ASCII;
         CTYPE H5T_C_S1;
      }
      DATASPACE  SCALAR
   }
   ATTRIBUTE "description" {
      DATATYPE  H5T_STRING {
         STRSIZE H5T_VARIABLE;
         STRPAD H5T_STR_NULLTERM;
         CSET H5T_CSET_ASCII;
         CTYPE H5T_C_S1;
      }
      DATASPACE  SCALAR
   }
   ATTRIBUTE "entries" {
      DATATYPE  H5T_STD_I64LE
      DATASPACE  SCALAR
```

*Source: As per TRIUMF RICHPID docs
https://github.com/TRIUMF-Capstone2022/RICHPID/blob/86a82e48ec32cad9fdf6bed99a57cfe8e1c0f024/docs/hdf5_format.txt

Data

Description: Large particle-tagged data sets for training

Data source: NA65 datasets generated from light sensing RICH detector (~2000 pixels) for 2016, 2017, 2018 and 2021
We would be using 2018 data (2 periods)

Data Format:
-HDF5 (File directory structured array data)

Number of training examples:
~ 11.5 Million from 2018 B and 2018 E periods

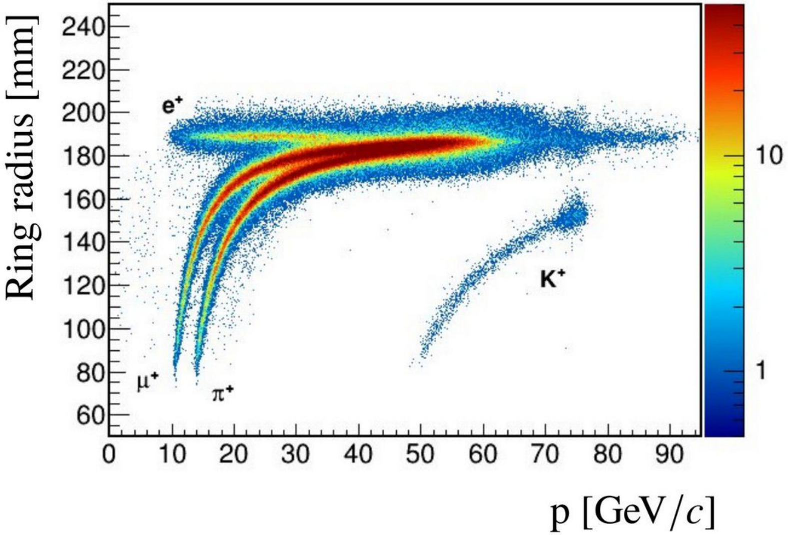
Class Imbalance

Particle * (Class)	**N***	**%**
Pion	**1,169,556**	**10 %**
Muon	10,281,301	89%
Positron	113,112	1 %
Total	11,563,969	

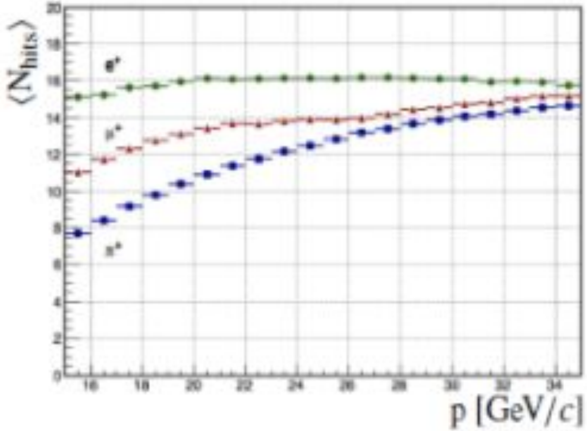
*Source: As per the initial datasets shared by TRIUMF for 2018 (B & E); Other 4 years data would be shared on need basis
https://github.com/TRIUMF-Capstone2022/RICHPID/tree/main/docs

Machine learning - key features

Ring radius vs. Momentum

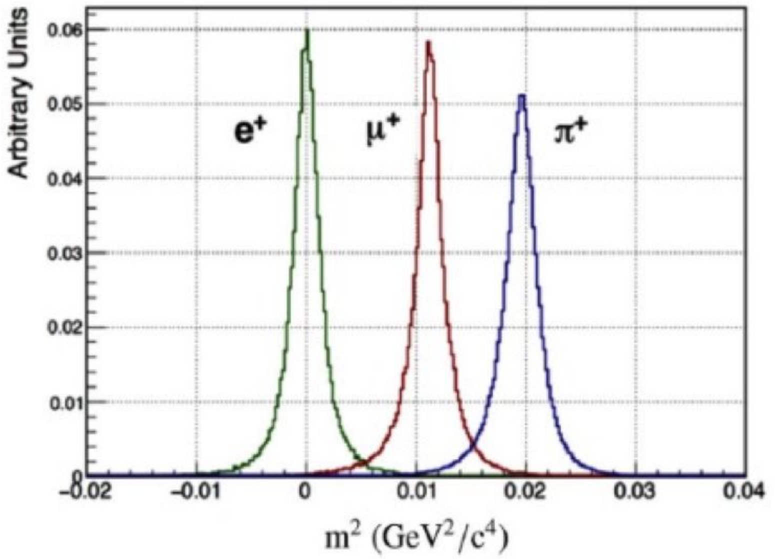


Hits vs. Momentum

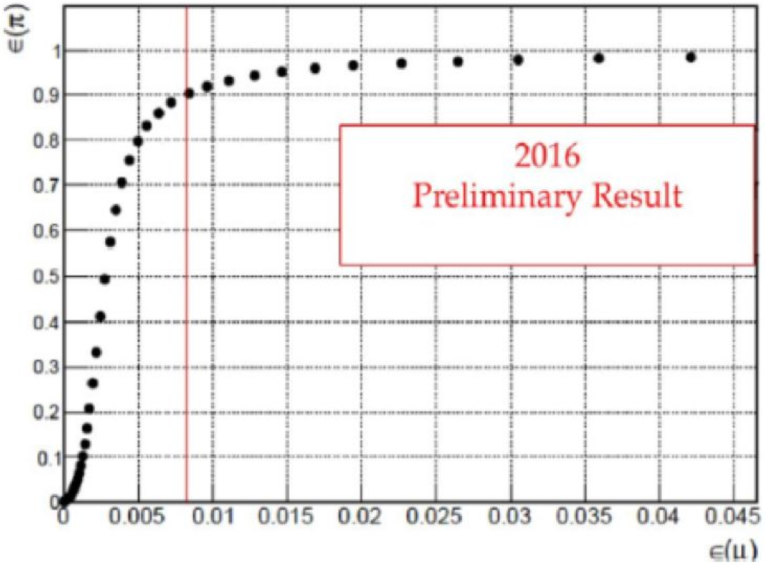


Machine learning - current analytical performance

Current Classification Results



Current ROC Curve


$\epsilon(\pi)$
$\epsilon(\mu)$
Preliminary Result