INFN
Istituto Nazionale di Fisica Nucleare

Institute of High Energy Physics
Chinese Academy of Sciences

LSTM Peak finding algorithm for Cluster Counting on Real Data and ML based Algorithm for Particle Identification

Muhammad Numan Anwar

Bari (INFN) - Beijing(IHEP)
Meeting

Feb 3 , 2026

- **Analysis of Real Test Beam Data**
- Configuration Setup for channel 5
- Optimization of LSTM Model by Search Grid
- Peak Finding Algorithm

- **ML based Algorithm on for Particle Identification(Ongoing)**
- Simulation Parameters
- Number of Primary Cluster (MC)

  I) Muon

  ii) Kaon
- NN (CNN) Reconstruction Results
- Particle Identification

Main Goal of the Presentation

Task 1: The first task is related to apply the best trained LSTM Model (Highest Auc value among all configurations) on 180 GeV muon real data in order to detect peaks in comparison with RTA Algorithm on the behalf of different selection cuts(0.2-0.3 0.4)

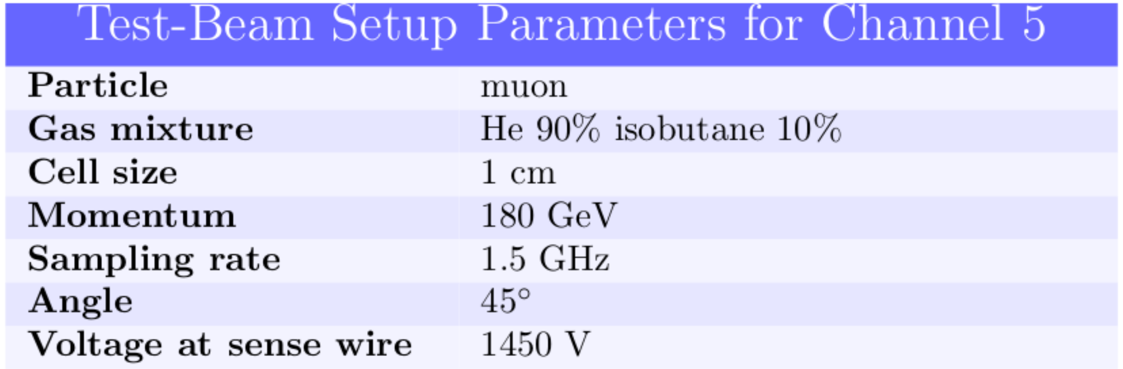
Task 2: The second task is related to particle identification of muon and kaon. That is why first we generated some simulated samples from 2-10 GeV momenta for Kaon. Then we apply best LSTM model to detect peaks in each waveform and then in the second step best CNN regression Model was applied to reconstruct the number of primary cluster on the behalf of detected peaks.

Note: I already assumed that all of you know about the two steps of cluster counting techniques. In the first step the best LSTM model find the peaks (Primary + Secondary electrons) in each waveform against the noise. While in the second step best CNN model estimate the number of primary clusters based on the detected peaks. That is why I showed the final results of Number of primary Clusters MC & CNN regression Model.

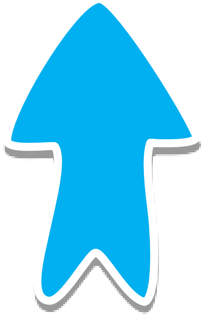
LSTM Peak Finding Algorithm on Real Data

Configuration Setup for the Real Test Beam data (180 GeV & Channel 5)

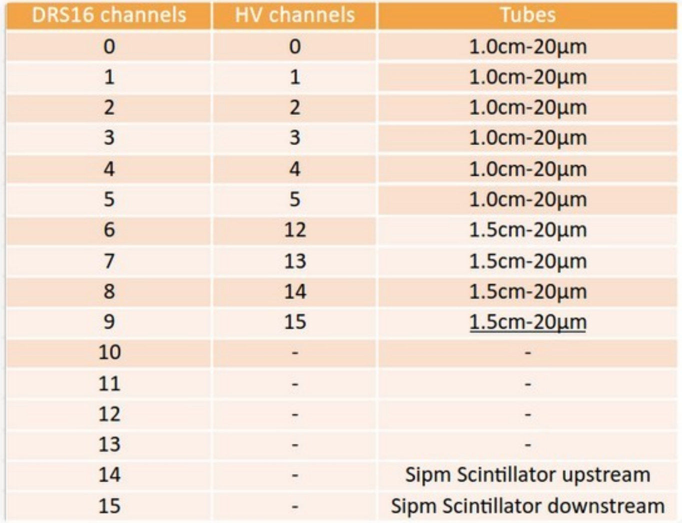
Confiurations and setup information



Test-Beam Setup Parameters for Channel 5	
Particle	muon
Gas mixture	He 90% isobutane 10%
Cell size	1 cm
Momentum	180 GeV
Sampling rate	1.5 GHz
Angle	45°
Voltage at sense wire	1450 V

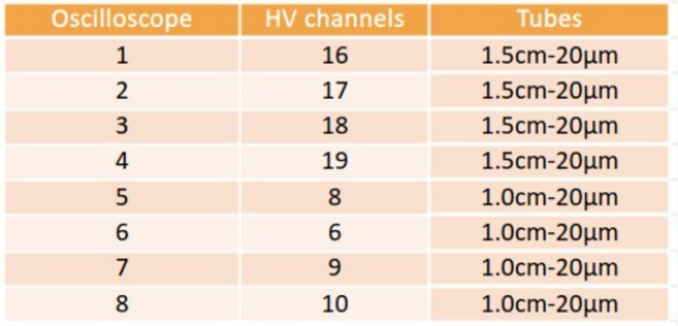


Channel 5 Setup Configuration



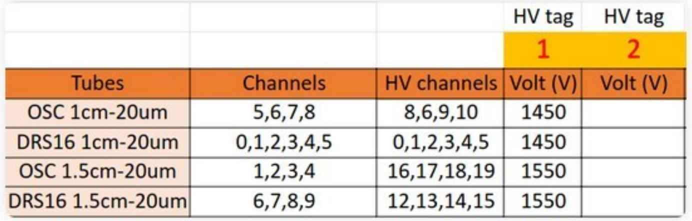
DRS16 channels	HV channels	Tubes
0	0	1.0cm-20μm
1	1	1.0cm-20μm
2	2	1.0cm-20μm
3	3	1.0cm-20μm
4	4	1.0cm-20μm
5	5	1.0cm-20μm
6	12	1.5cm-20μm
7	13	1.5cm-20μm
8	14	1.5cm-20μm
9	15	1.5cm-20μm
10	-	-
11	-	-
12	-	-
13	-	-
14	-	Sipm Scintillator upstream
15	-	Sipm Scintillator downstream

Tubes setup with DRS



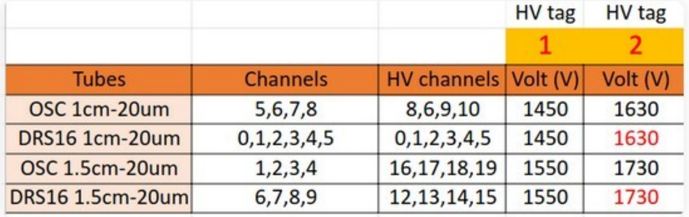
Oscilloscope	HV channels	Tubes
1	16	1.5cm-20μm
2	17	1.5cm-20μm
3	18	1.5cm-20μm
4	19	1.5cm-20μm
5	8	1.0cm-20μm
6	6	1.0cm-20μm
7	9	1.0cm-20μm
8	10	1.0cm-20μm

Tubes setup with OSC



			HV tag	HV tag
			1	2
Tubes	Channels	HV channels	Volt (V)	Volt (V)
OSC 1cm-20um	5,6,7,8	8,6,9,10	1450	
DRS16 1cm-20um	0,1,2,3,4,5	0,1,2,3,4,5	1450	
OSC 1.5cm-20um	1,2,3,4	16,17,18,19	1550	
DRS16 1.5cm-20um	6,7,8,9	12,13,14,15	1550	

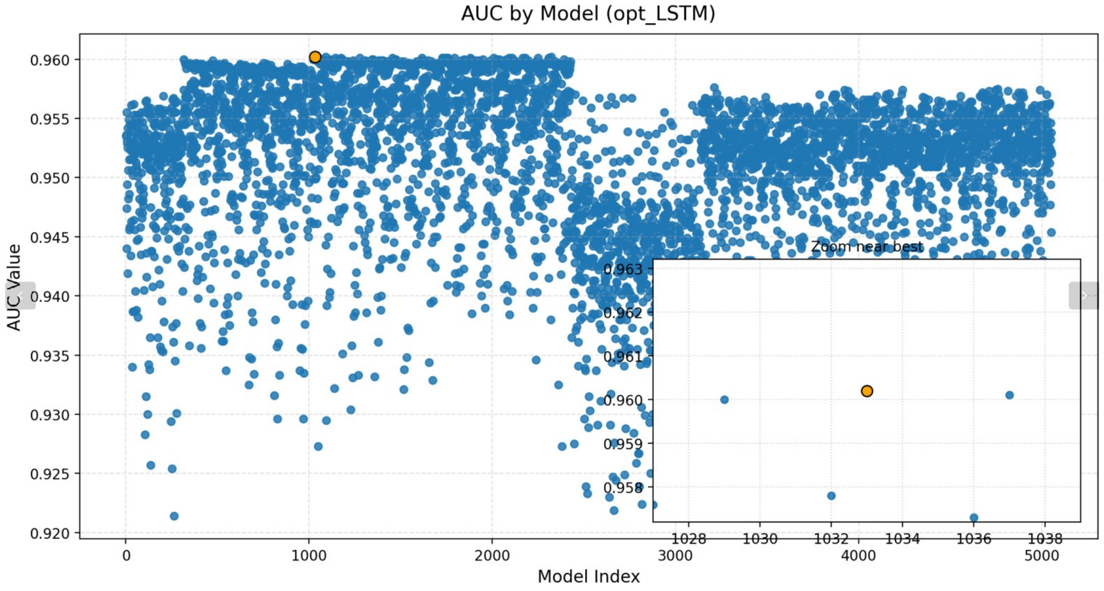
HV tags with 90/10 gas mix runs

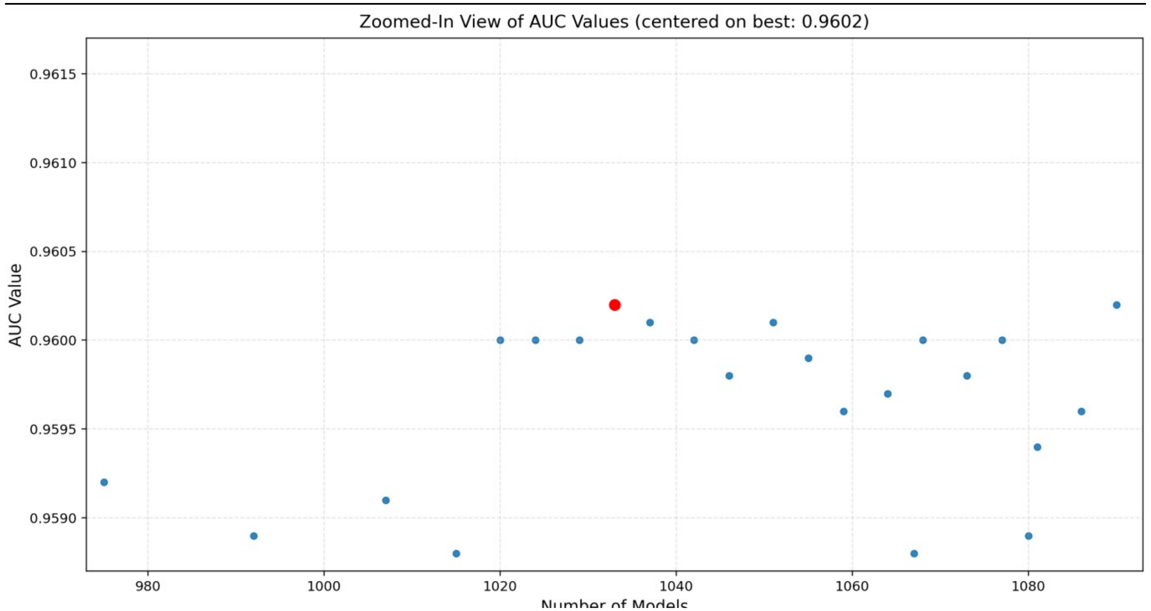


			HV tag	HV tag
			1	2
Tubes	Channels	HV channels	Volt (V)	Volt (V)
OSC 1cm-20um	5,6,7,8	8,6,9,10	1450	1630
DRS16 1cm-20um	0,1,2,3,4,5	0,1,2,3,4,5	1450	1630
OSC 1.5cm-20um	1,2,3,4	16,17,18,19	1550	1730
DRS16 1.5cm-20um	6,7,8,9	12,13,14,15	1550	1730

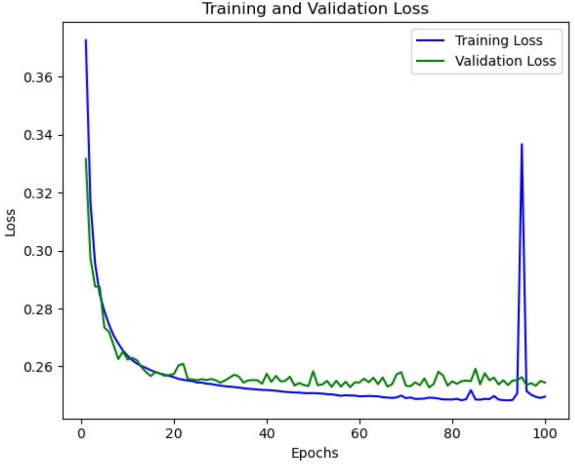
HV tags with 85/15 gas mix runs

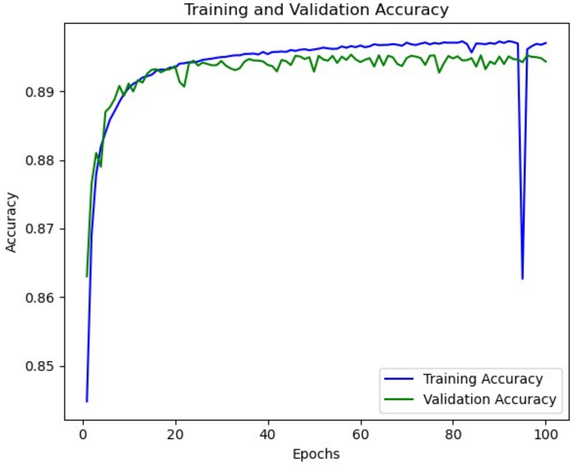
Search Grid Optimzation: BEST LSTM MODEL ON TUNE SIMULATED WAVEFORMS (50000)

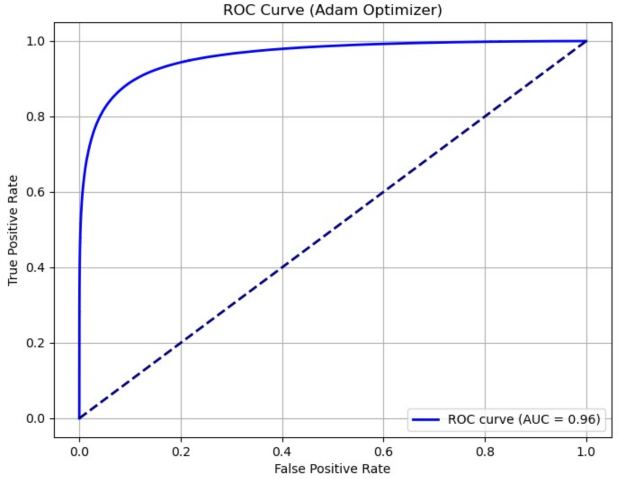


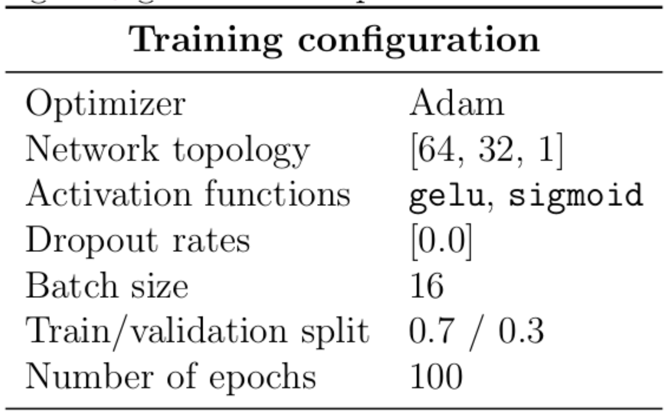


- We used different sets of hyper-parameters like activation functions, optimizers etc to train LSTM peak finding model on 50000 tuned simulated waveforms on the simulation parameter(showed in the previous table) and then we select the best model with highest area under the curve value (0.96) among all configurations showed by red dot in left Zoom view of AUC value.





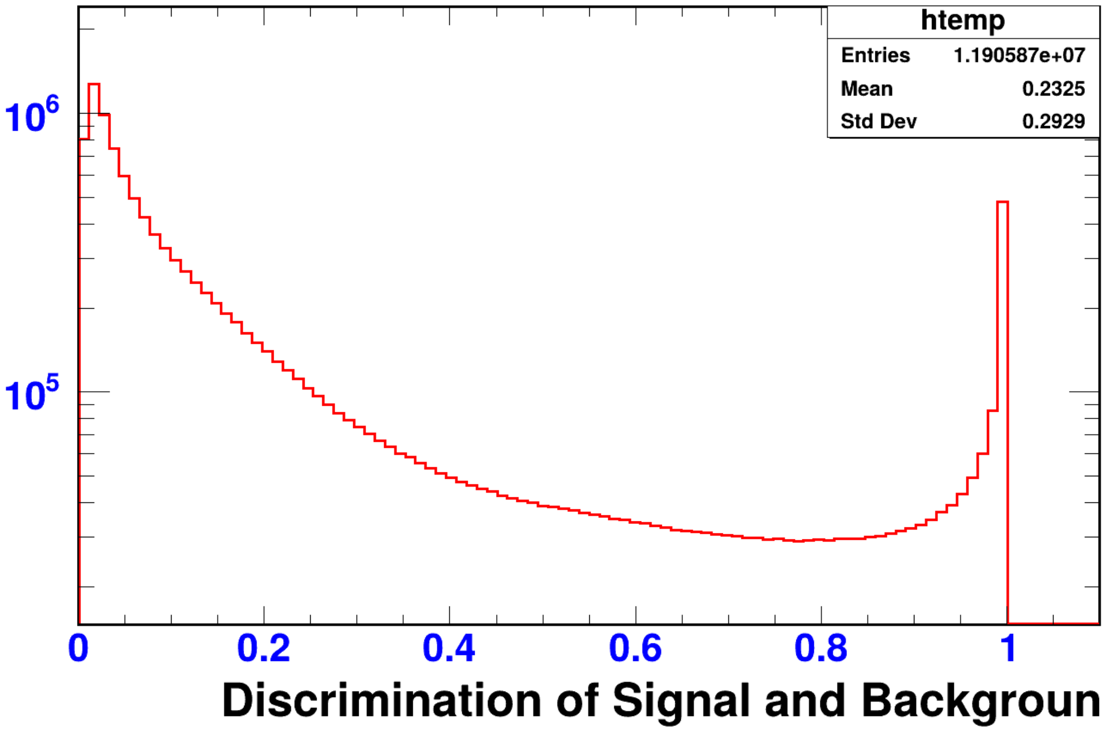


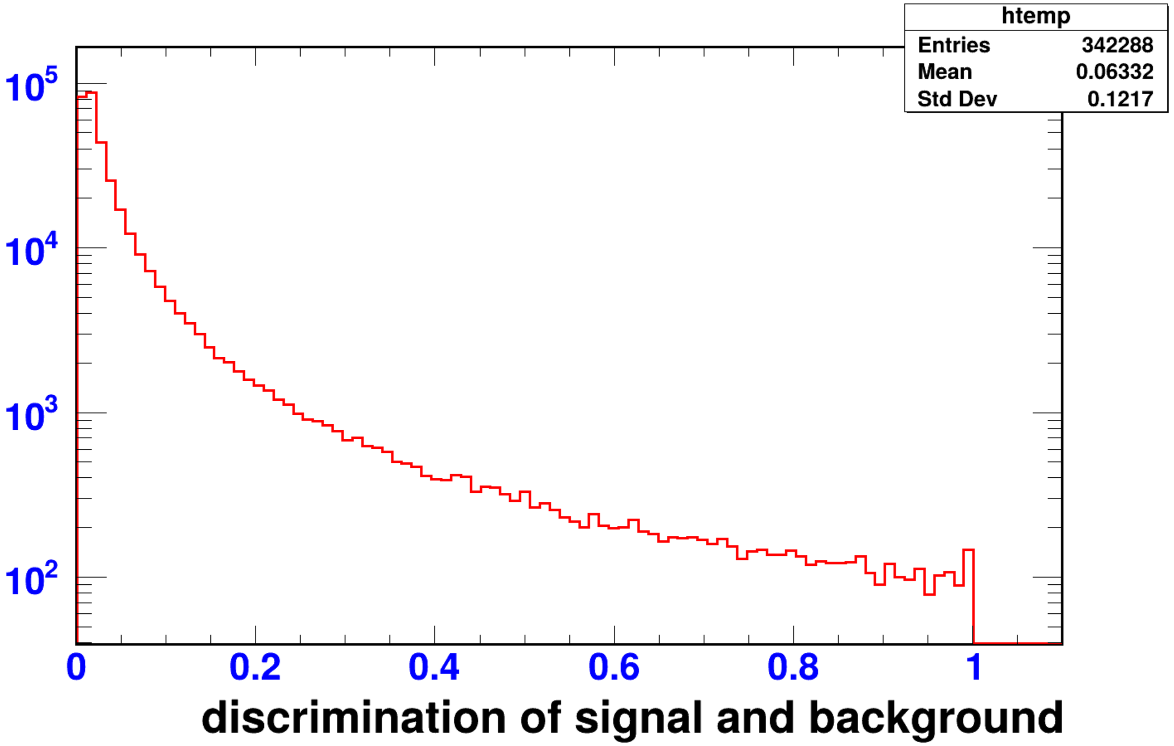


Training configuration	
Optimizer	Adam
Network topology	[64, 32, 1]
Activation functions	gelu, sigmoid
Dropout rates	[0.0]
Batch size	16
Train/validation split	0.7 / 0.3
Number of epochs	100

Hyperparameters of Best LSTM Model

TUNE SIMULATED VS Data Testing Part by using Best LSTM Model (AUC =0.96)



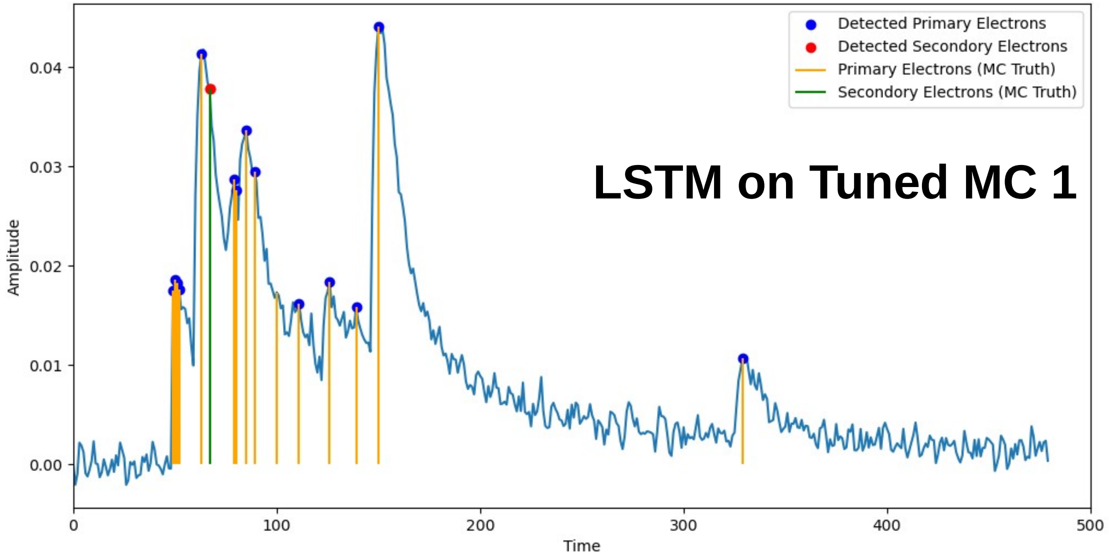


We applied best LSTM model on tune MC samples to discriminate signal and Background

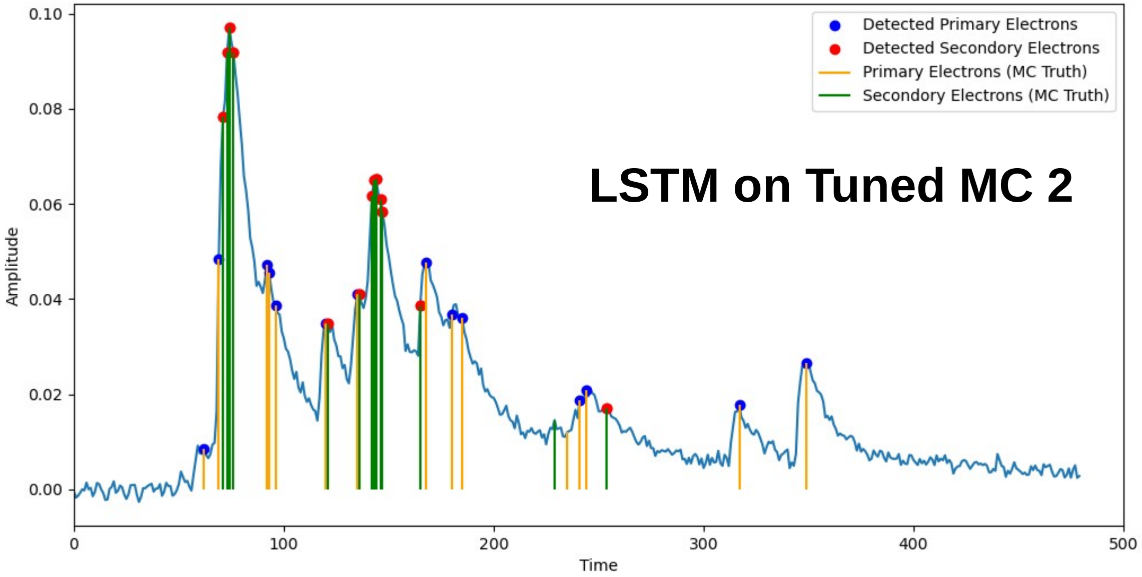
We applied best LSTM model on data (0.3) to discriminate signal and BackgroundData with Log Scale

LSTM Peak Finding Waveforms Results of MC & Data In Comparison with RTA/Derivative Algorithm

Monte Carlo (MC)

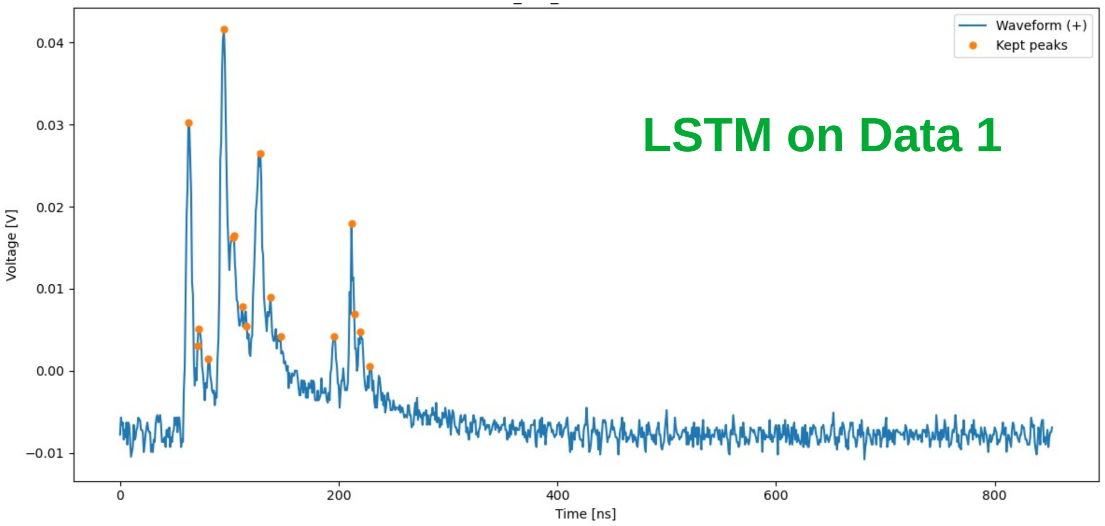


LSTM on Tuned MC 1

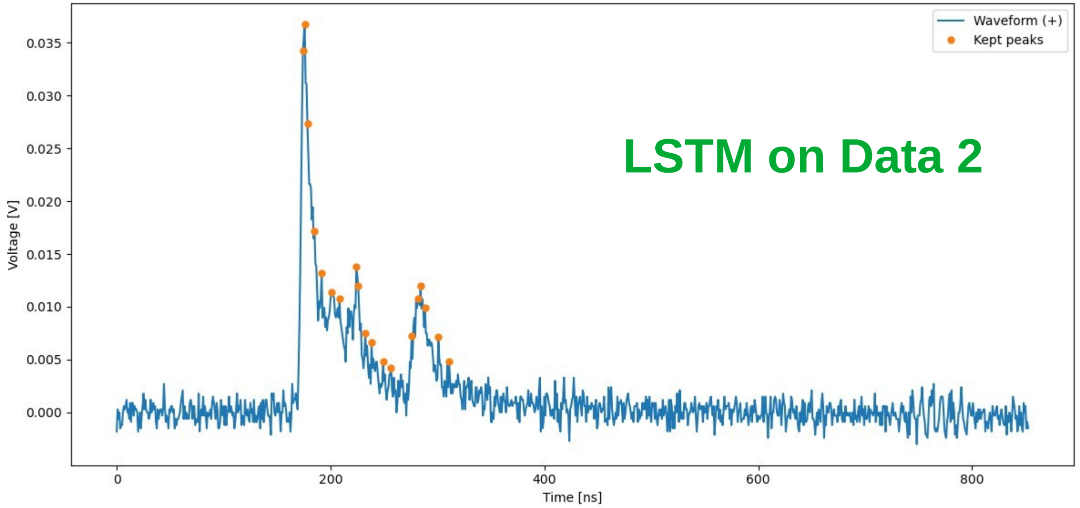


LSTM on Tuned MC 2

180 GeV Real Test

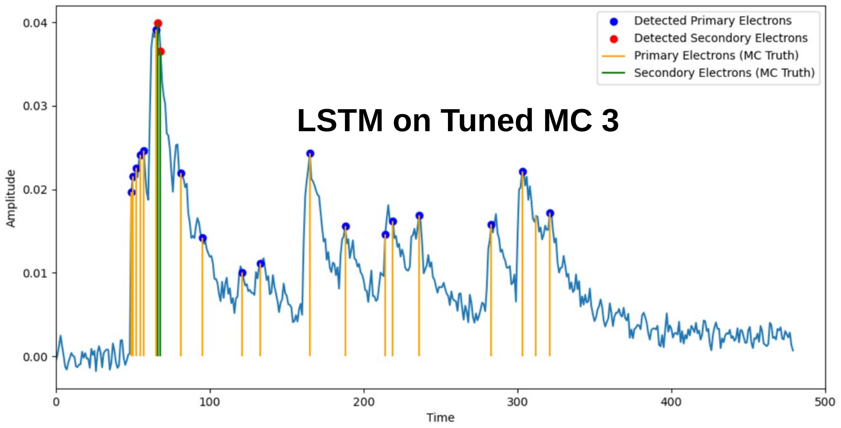


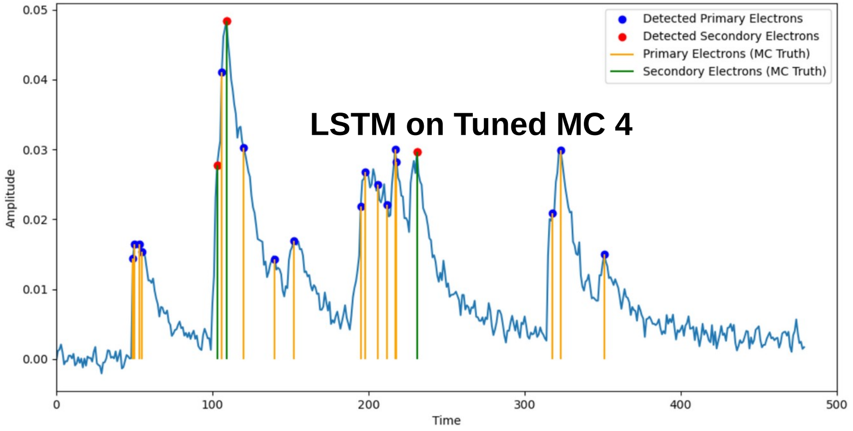
LSTM on Data 1

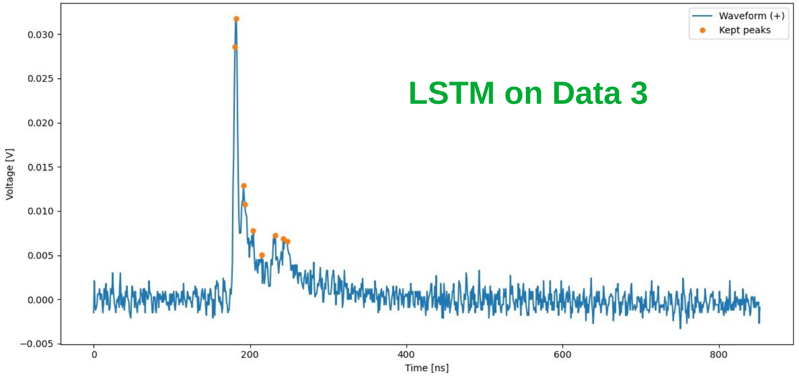


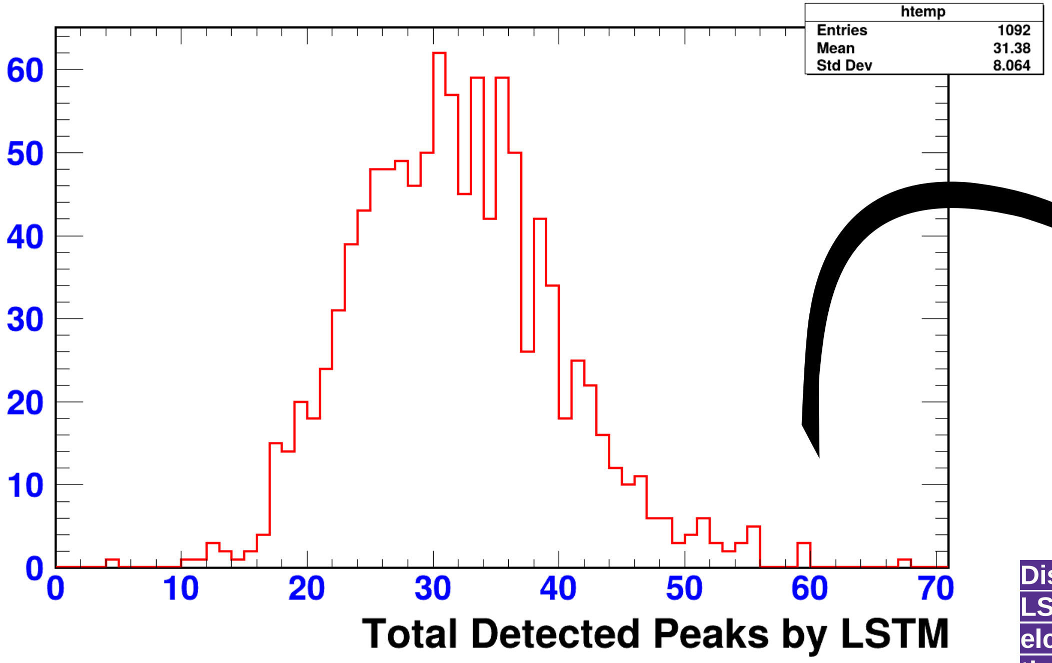
LSTM on Data 2

LSTM Peak Finding Results of MC & Data In Comparison with RTA/Derivative Algorithm

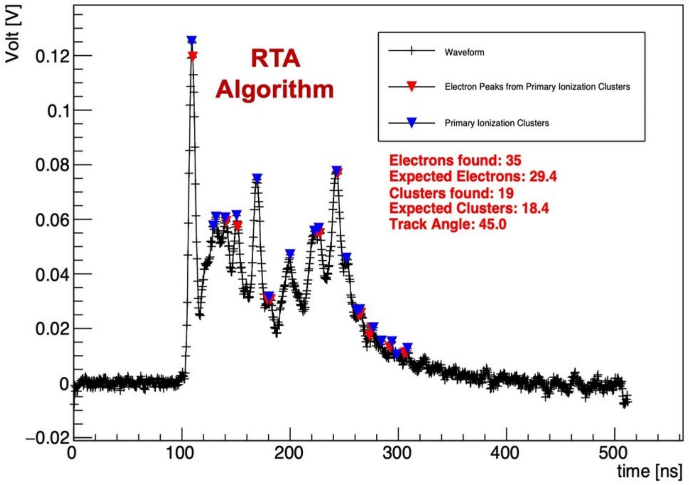

LSTM on Tuned MC 3


LSTM on Tuned MC 4

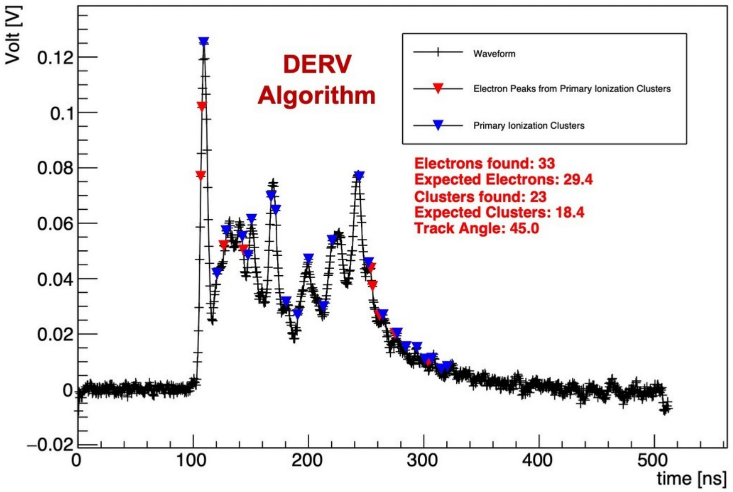

LSTM on Data 3



Distribution of total detected peaks using LSTM arises from primary and secondory elctrons (31.38) for channel 5 data with threshould 0.3

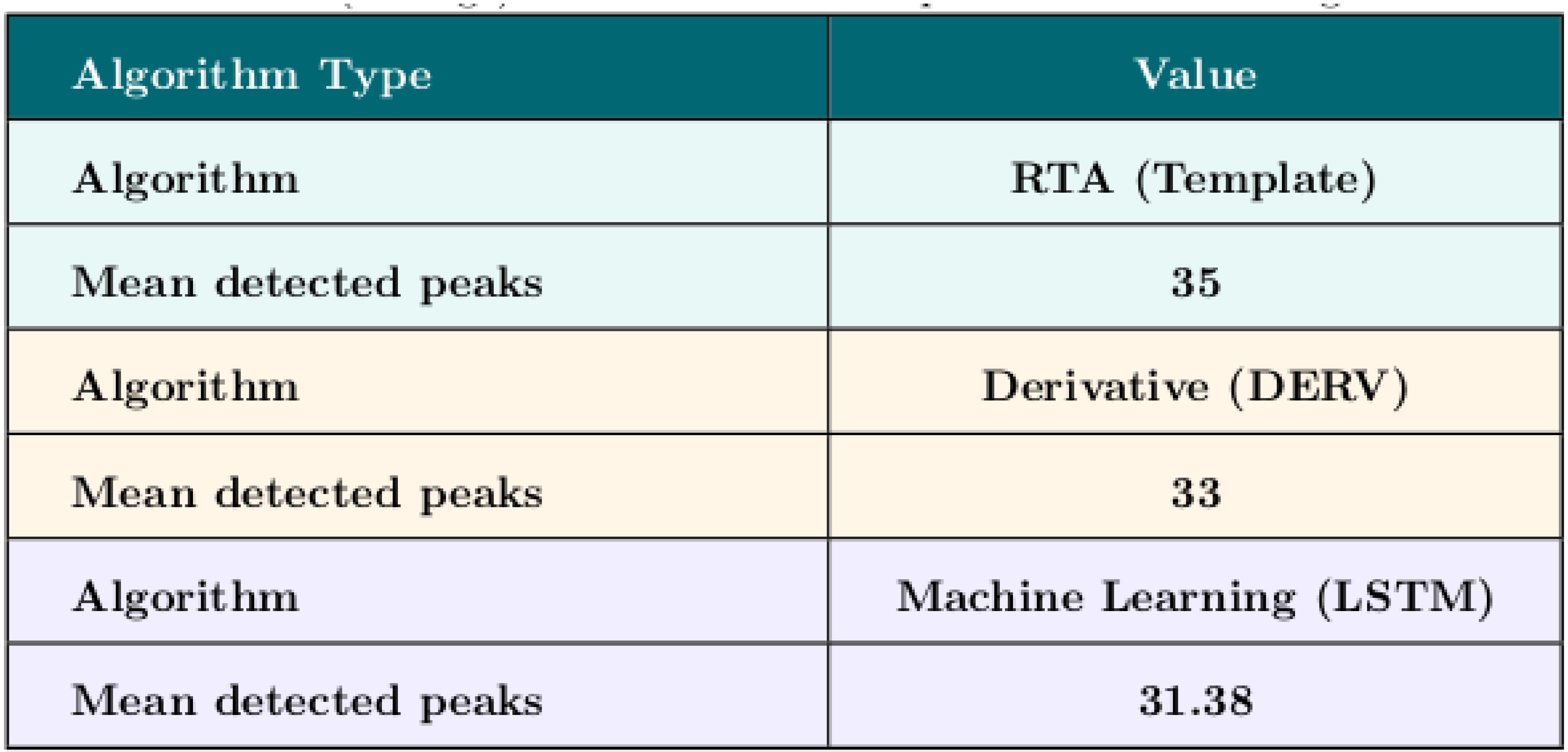

RTA Algorithm

Electrons found: 35
Expected Electrons: 29.4
Clusters found: 19
Expected Clusters: 18.4
Track Angle: 45.0


DERV Algorithm

Electrons found: 33
Expected Electrons: 29.4
Clusters found: 23
Expected Clusters: 18.4
Track Angle: 45.0

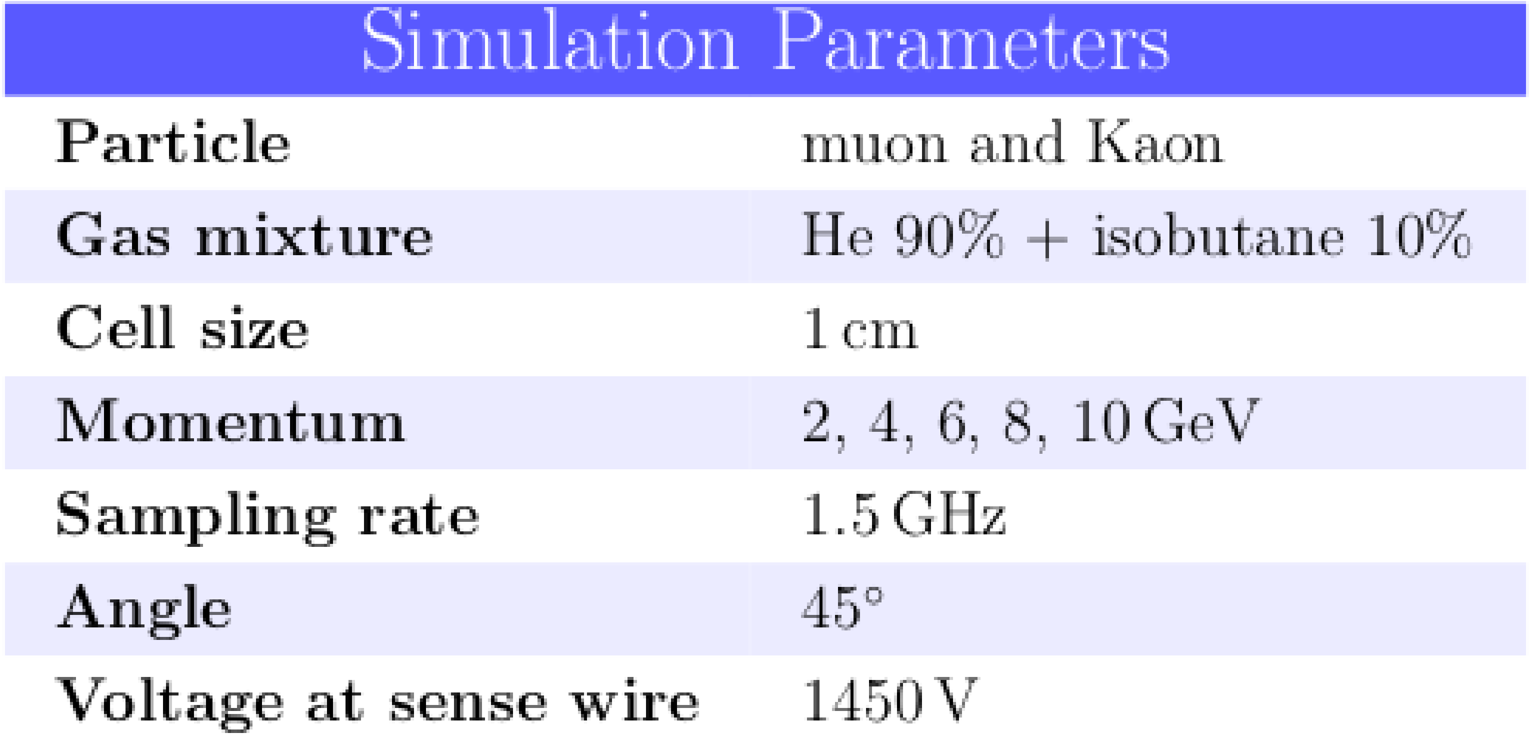
Conclusion



Algorithm Type	Value
Algorithm	RTA (Template)
Mean detected peaks	35
Algorithm	Derivative (DERV)
Mean detected peaks	33
Algorithm	Machine Learning (LSTM)
Mean detected peaks	31.38

ML Based Algorithm for ParticleIdentification(Simulated samples + NN) (Ongoing)

ML Based Algorithm for Particle Identification(Ongoing): Simulation Parameters

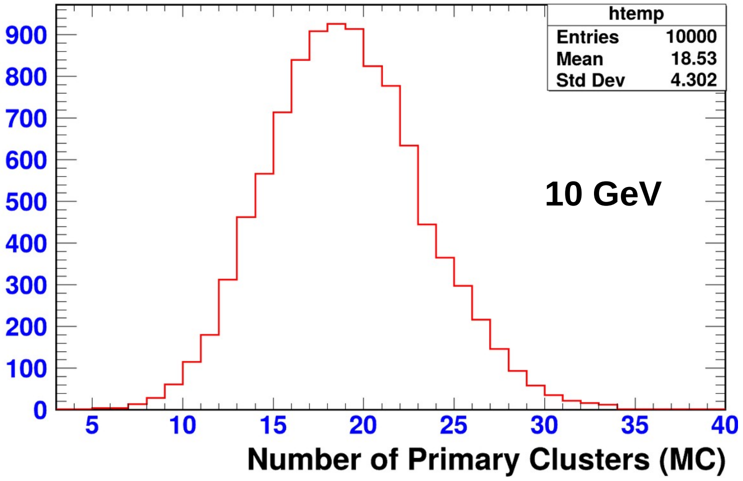


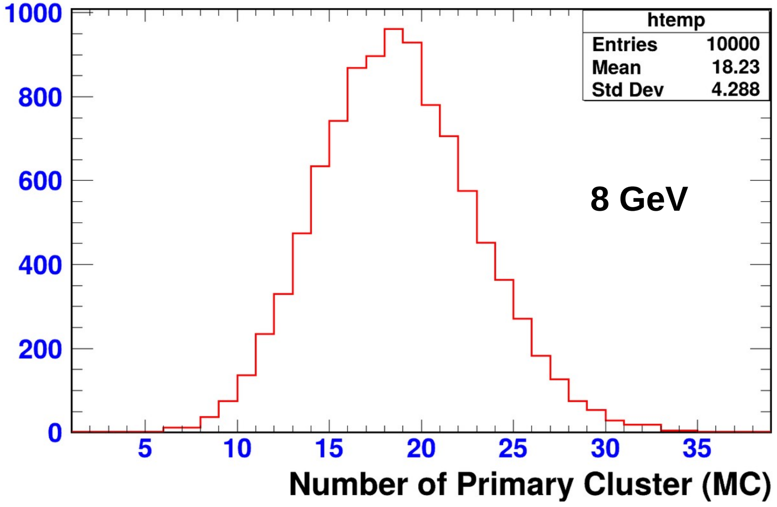
Simulation Parameters	
Particle	muon and Kaon
Gas mixture	He 90% + isobutane 10%
Cell size	1 cm
Momentum	2, 4, 6, 8, 10 GeV
Sampling rate	1.5 GHz
Angle	45°
Voltage at sense wire	1450 V

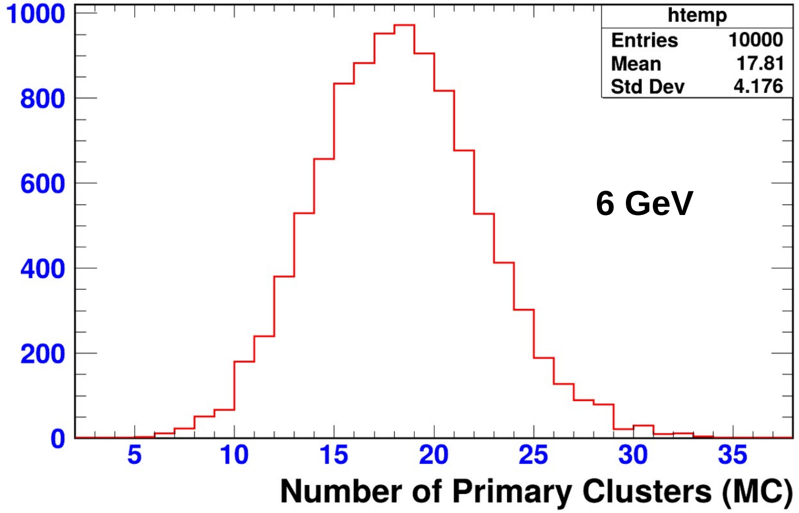
Above are Simulation parameters were used which matched the real test beam data.

Number of Primary Cluster (MC)

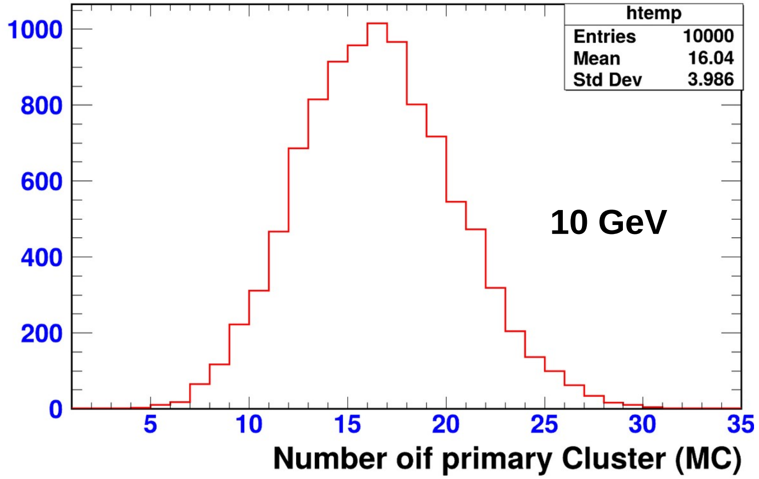
For Muon

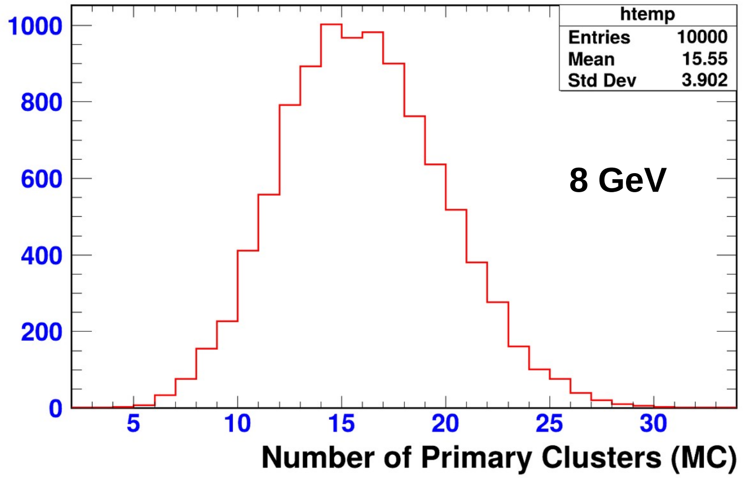


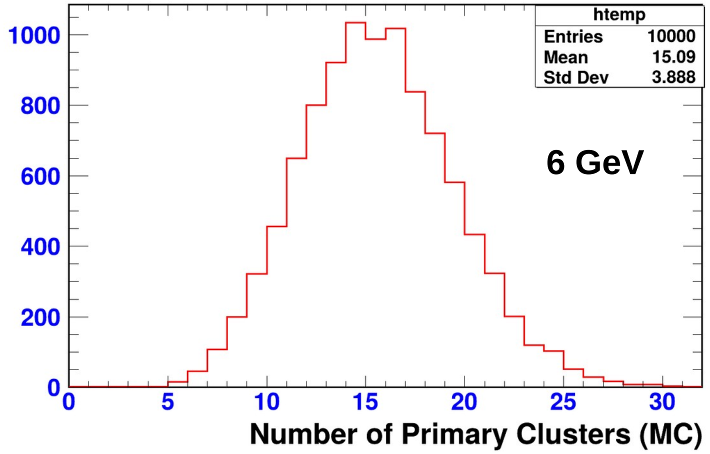




For Kaon

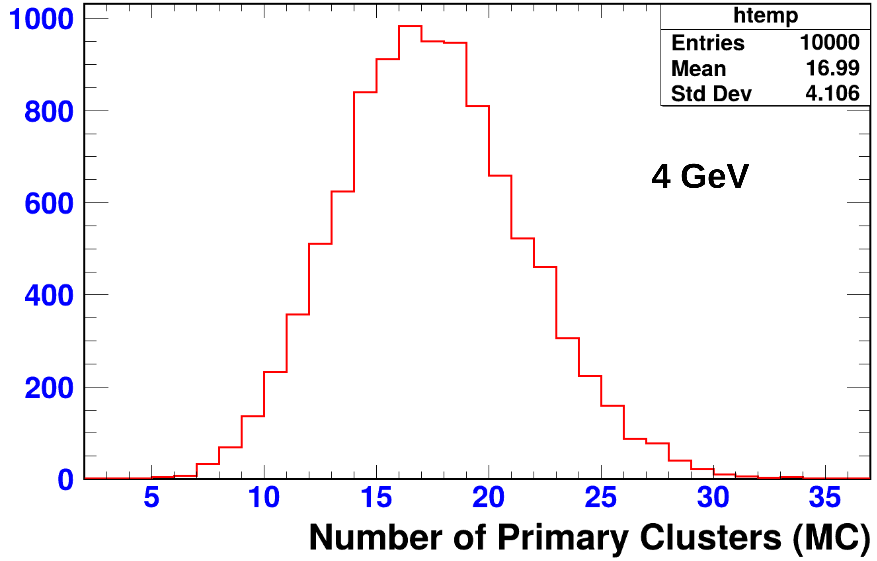


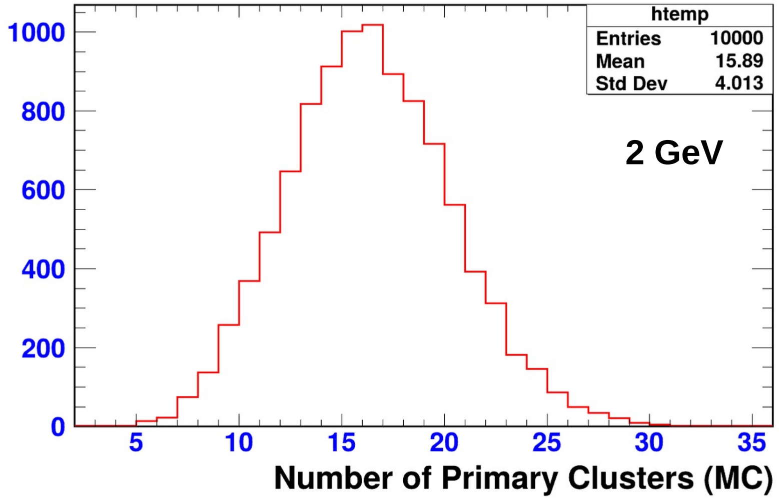




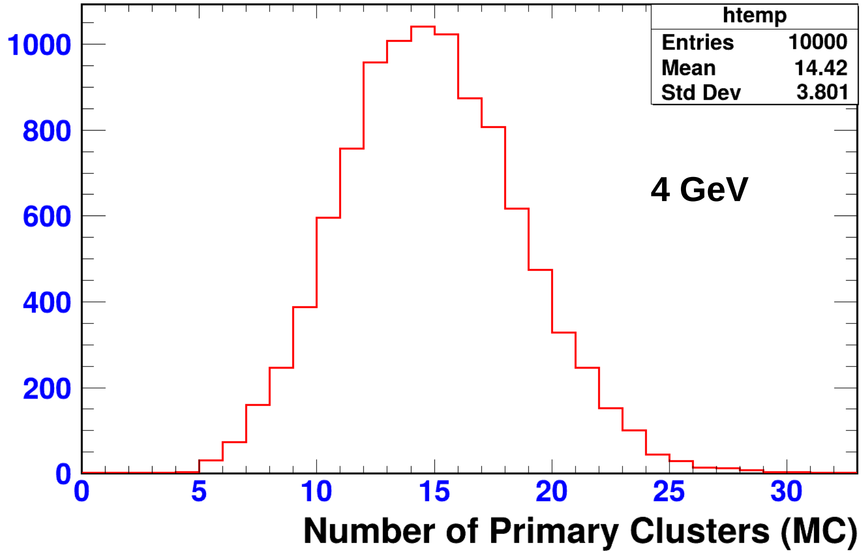
Number of Primary Cluster (MC)

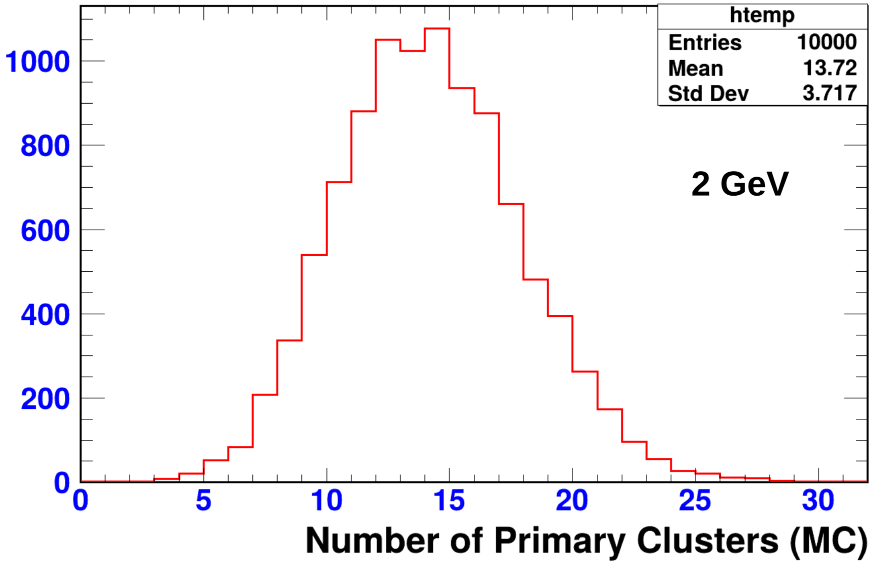
For Muon





For Kaon

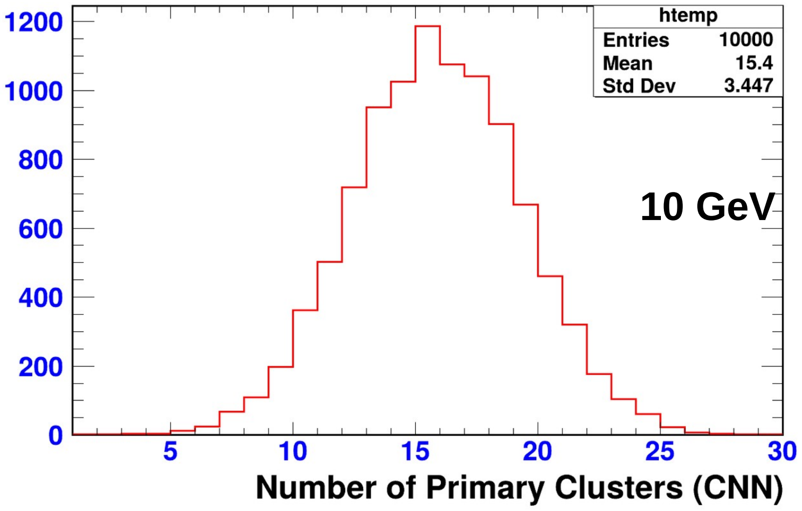


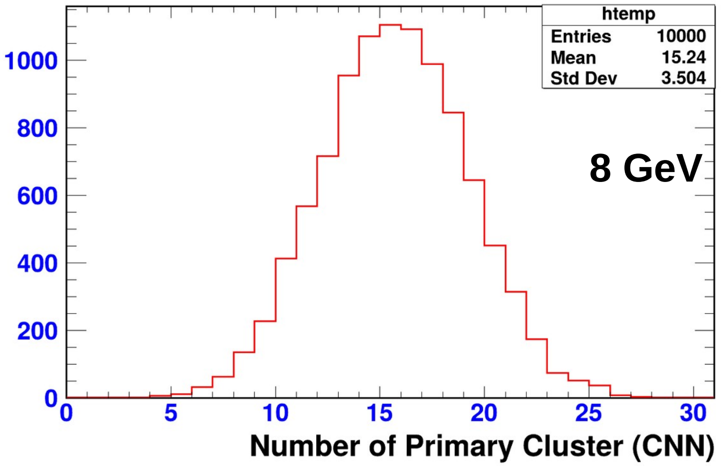


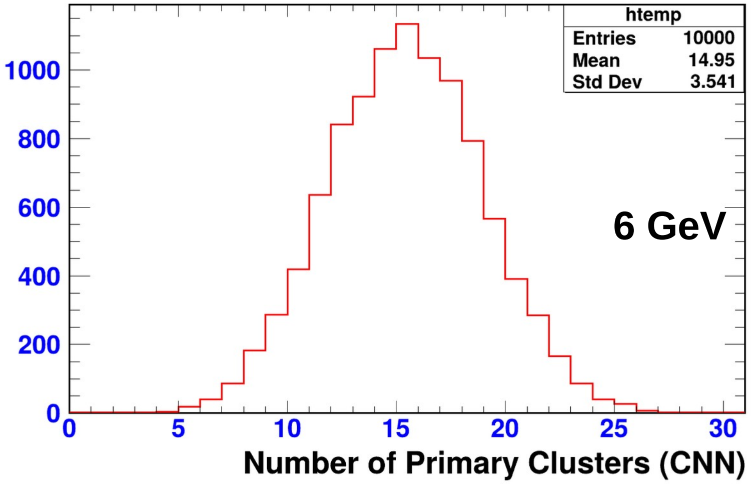
All these distributions are related to the number of primary clusters Monet carlo (MC) for the momentum (2-10 Gev) of muon and Kaon. We also need to generated it for 180 GeV momentum

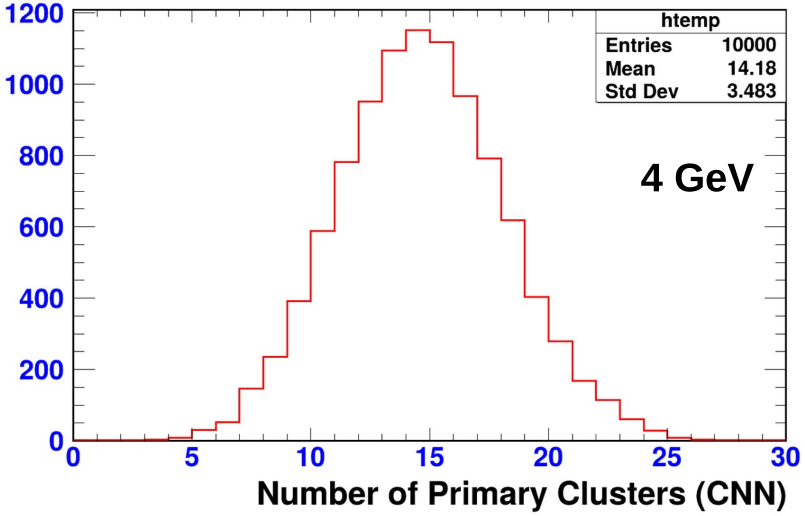
Reconstruction of Number of Primary Clusters (CNN)

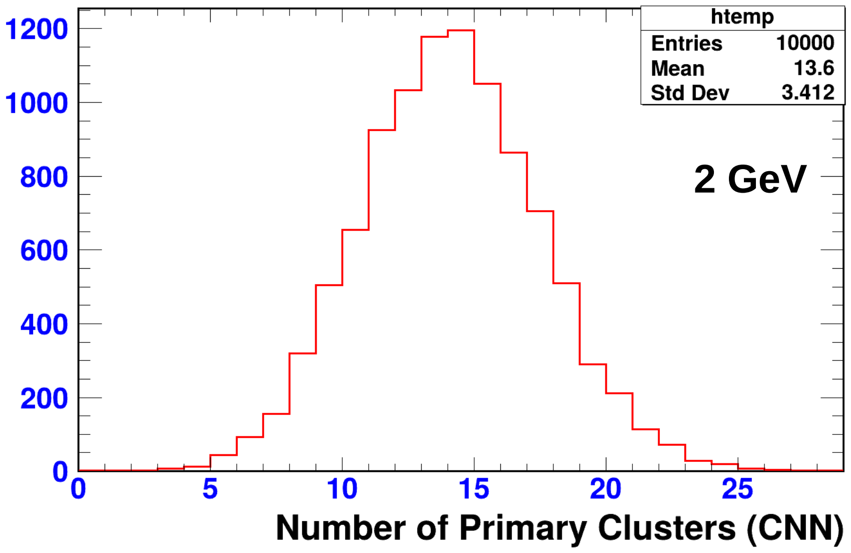
For Kaon







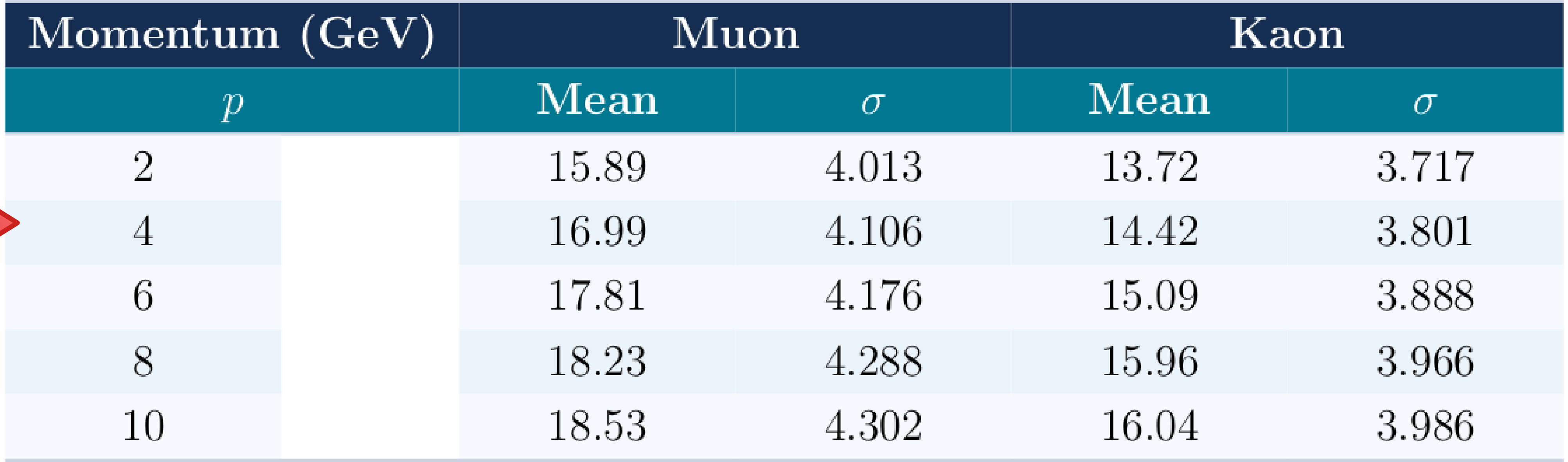




All these distributions are related to the number of primary clusters reconstructed by CNN Model for the momentum (2-10 Gev) of Kaon

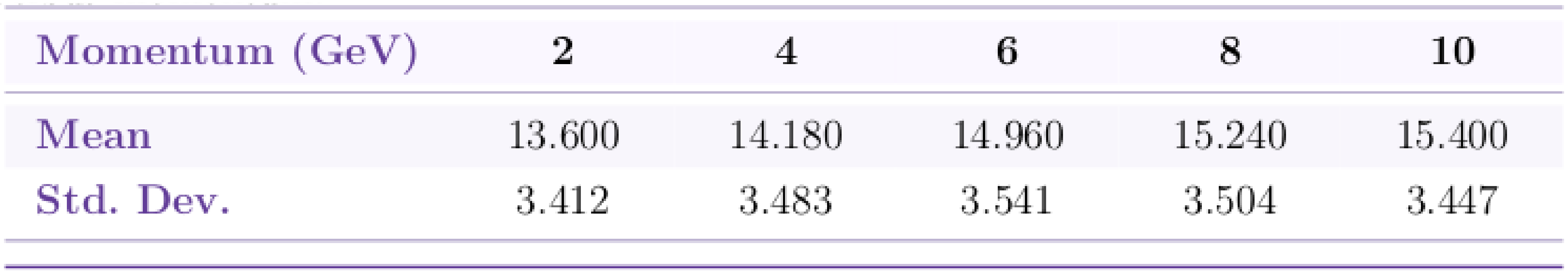
MC truth and Final Results of Reconstructions for Kaon By CNN Clusterization Model

Number of primary Clusters MC Truth (Muon and Kaon)



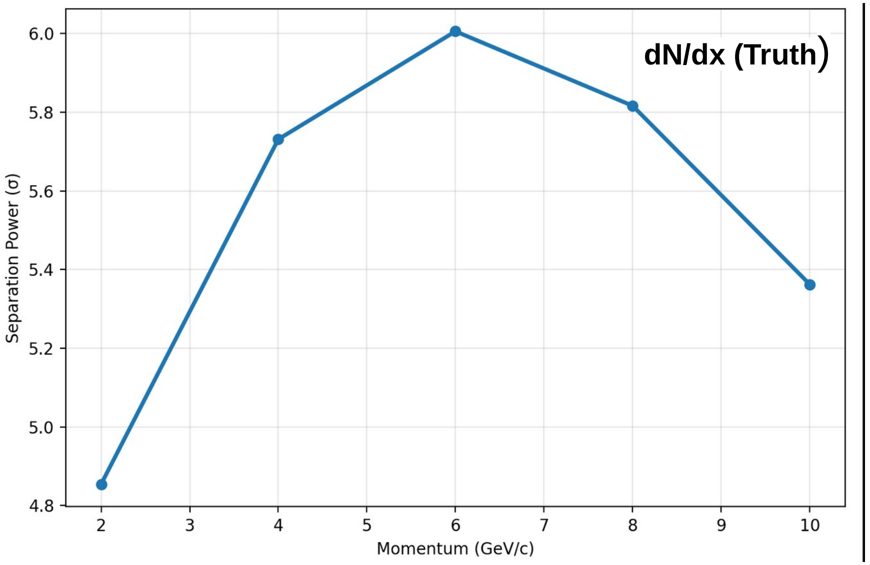
Momentum (GeV)	Muon		Kaon	
p	Mean	σ	Mean	σ
2	15.89	4.013	13.72	3.717
4	16.99	4.106	14.42	3.801
6	17.81	4.176	15.09	3.888
8	18.23	4.288	15.96	3.966
10	18.53	4.302	16.04	3.986

CNN Clusterization Results for Kaon (2-10 GeV). The mean number of primary clusters (CNN) matches with the MC



Momentum (GeV)	2	4	6	8	10
Mean	13.600	14.180	14.960	15.240	15.400
Std. Dev.	3.412	3.483	3.541	3.504	3.447

Super Preliminary and Incomplete PID performances for MC and NN Reconstruction



$$S = \frac{|\mu_\mu - \mu_K|}{(\sigma_\mu + \sigma_K)/2}$$

The above plot is just the seperation power for MC truth of 2, 4, 6, 8 and 10 GeV momenta of muon and kaon while for the NN reconstruction, I will do soon

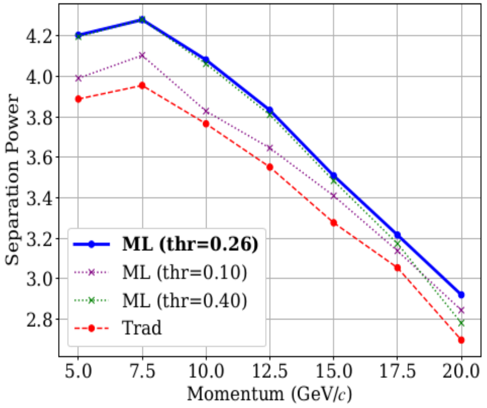


Fig. 9. The K/π separation power as a function of track momentum for 1 m track length. The red dashed line is from the traditional algorithm. The blue solid, violet dotted and green dotted lines are from the ML-based algorithm with a threshold of 0.26, 0.10 and 0.40, respectively. The blue solid line with a threshold of 0.26 achieves the overall best performance, which has a roughly 10% better K/π separation than the traditional algorithm.

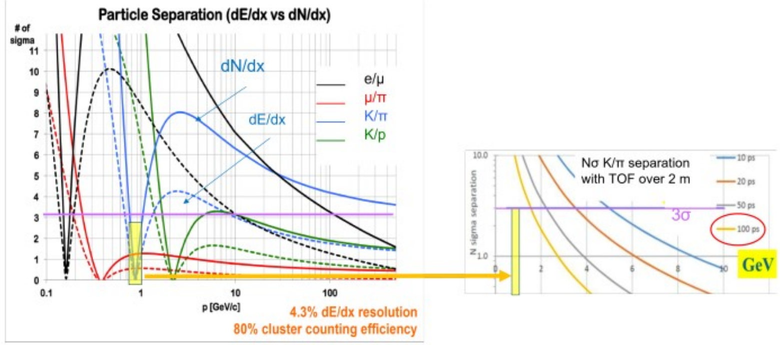


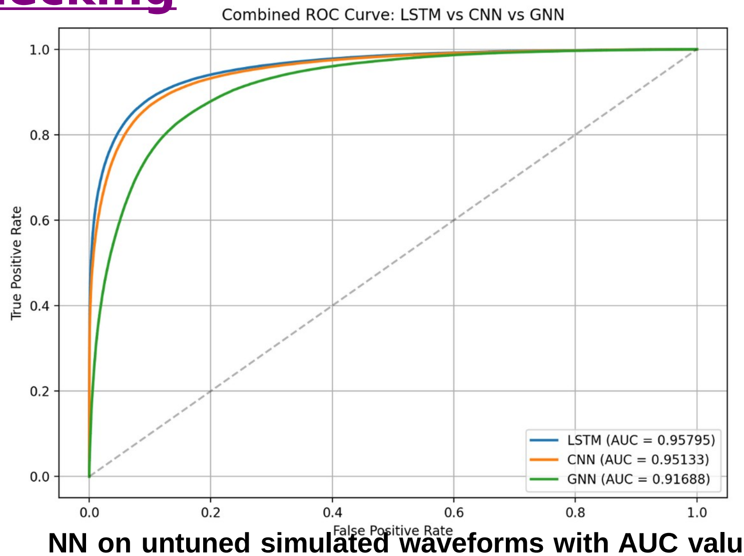
Figure 1: Left: Analytic evaluation of particle separation capabilities achievable with dE/dx (solid curves) and dN/dx (dashed curves). The region between 0.85 GeV/c and 1.05 GeV/c where a different technique is needed is highlighted in yellow. Right: PID performance as a function of the time resolution by using a time of flight technique over 2 m to recover the particle identification around 1 GeV.

What to do Next

- Now we will train CNN clusterization model and apply the model on MC and data to estimate the number of primary clusters based on the detected peaks
- All the results should be compare with RTA algorithm
- ML based Algorithm for PID should also be completed for muon and kaon

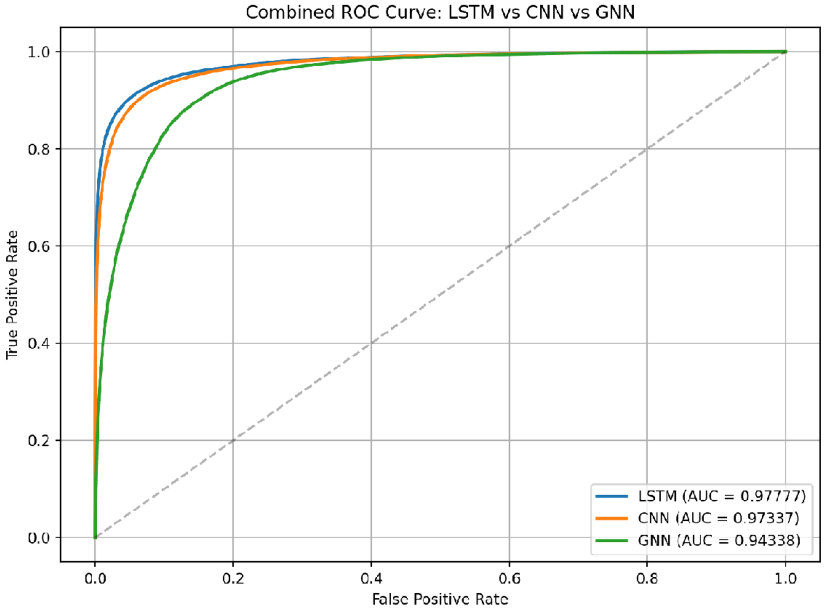
Models optimzation Just for Checking Which I did. Best is LSTM

AUC Value: LSTM VS CNN VS GNN for 180 GeV Tune Simulated waveforms (50000)



NN on untuned simulated waveforms with AUC value. It is also done for other three models too like DGCNN but the best one I found LSTM everywhere

AUC Value: LSTM VS CNN VS GNN for 180 GeV UnTune Simulated waveforms (50000)



NN on tuned simulated waveforms with AUC value. It is also done for other three models too like DGCNN but the best one I found LSTM everywhere

Thank
you

1) THESIS (Priority)

January 19 – February 5

➔ **Complete the following thesis chapters:**

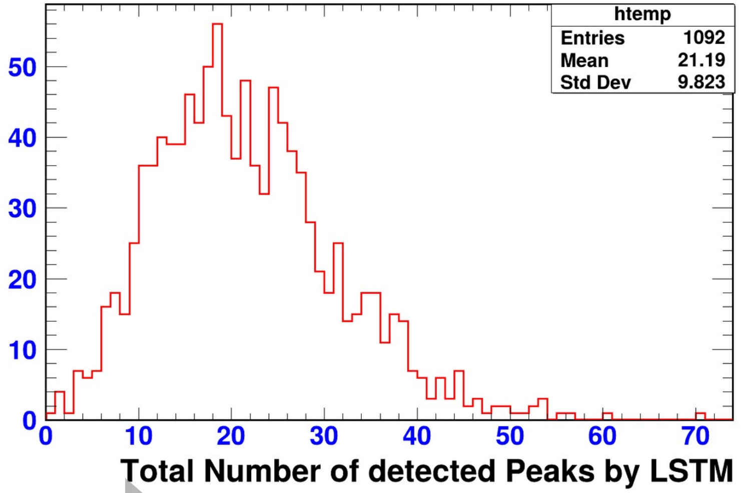
- Introduction to the ICSC Project (Completed)
- Introduction to FCC and IDEA Detector (Completed)
- Full Simulation of the IDEA Drift Tube Using Garfield++ (Completed)
- Deep Neural Network Models in Machine Learning (45 % completed)
- Machine-Learning-Based Cluster Counting for Particle Identification

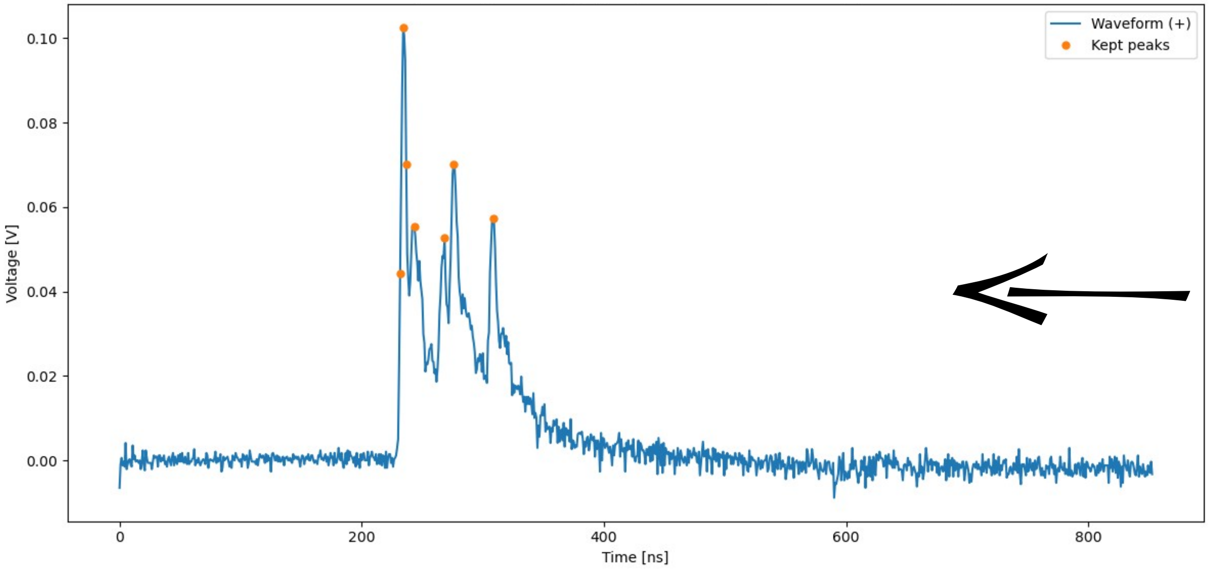
February 6 – February 28

➔ **Complete the remaining two thesis chapters:**

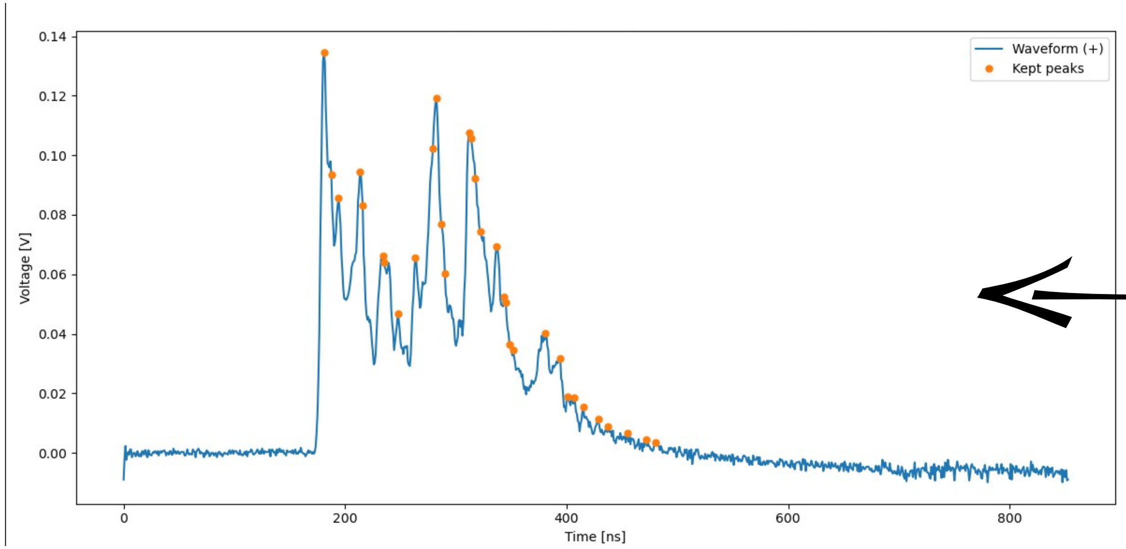
- Hyperparameter Optimization Using High-Performance Computing Resources and Final Reconstructed Results of Simulations and Neural Networks

Selection cuts on real data (Selection cut /Threshold)



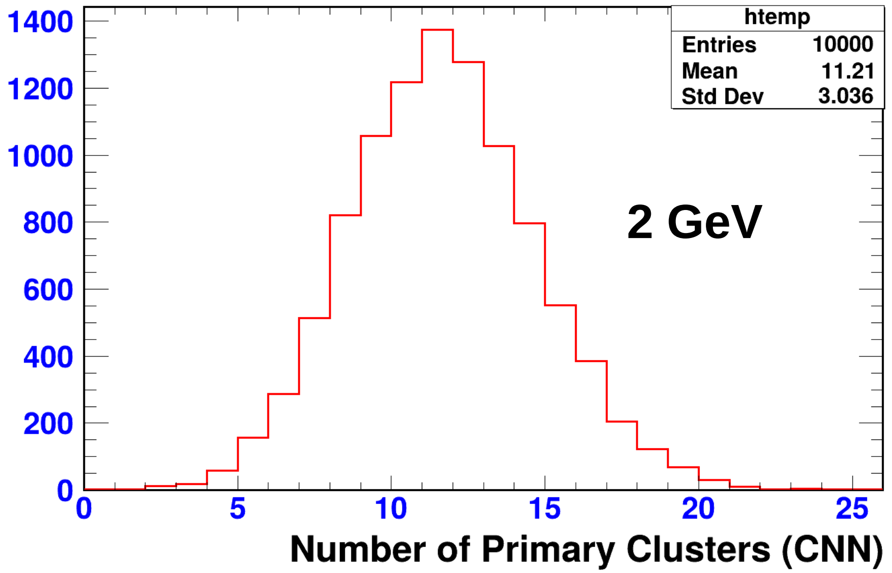


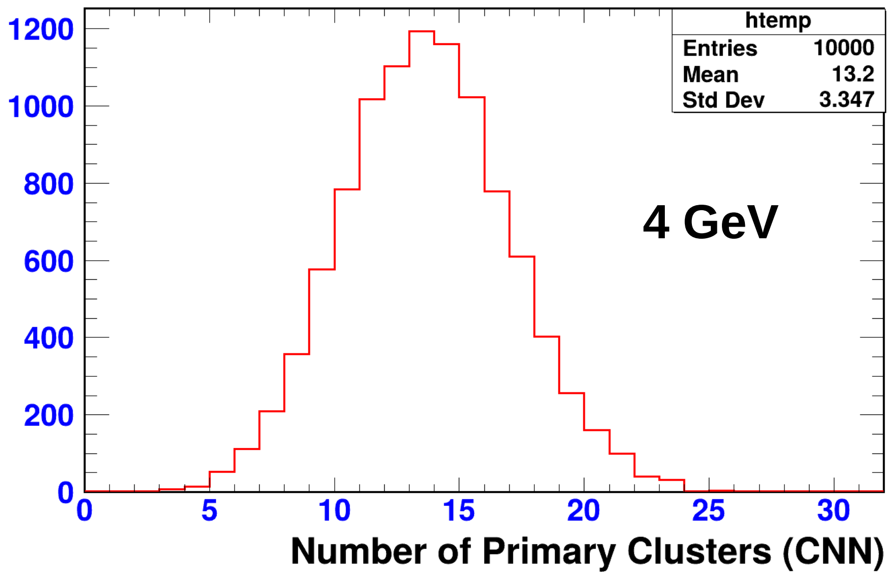
At Selection cut/Threshold = 0.4, less peaks appear as result found less mean value of the detected peaks in the distribution

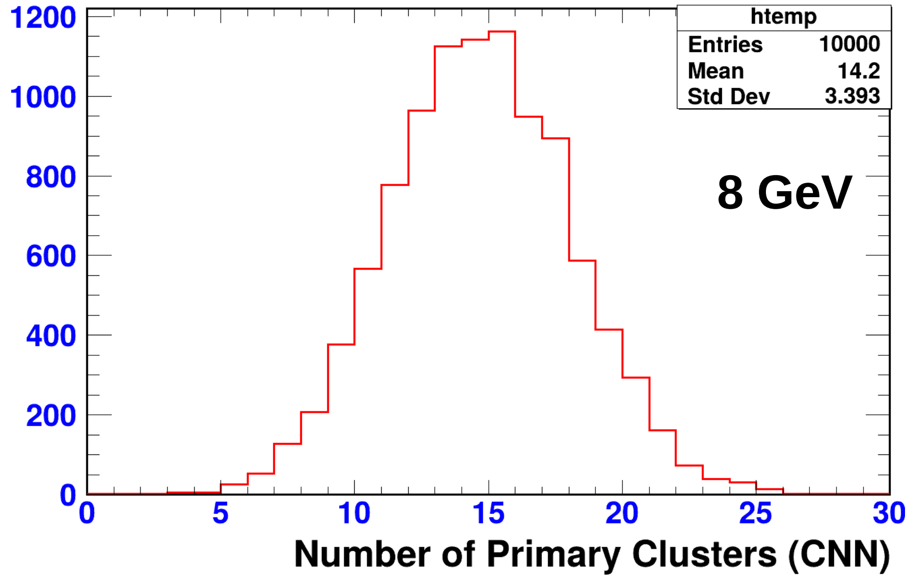


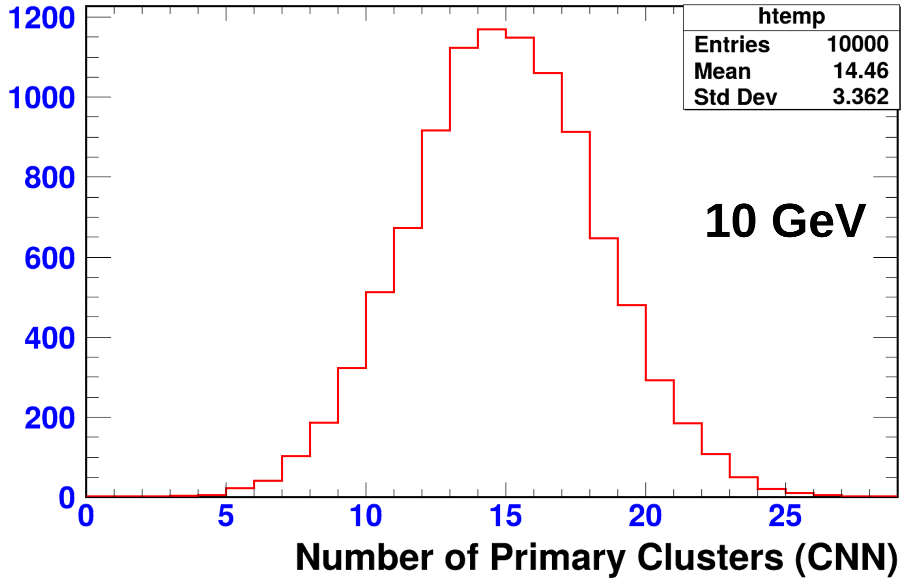
At Threshold = 0.25, some peak appear in the region of noise amplitude peaks appear as result found some fake peaks in mean value of the detected peaks in the distribution

Selection cut (0.9) on Simulated samples of Kaon





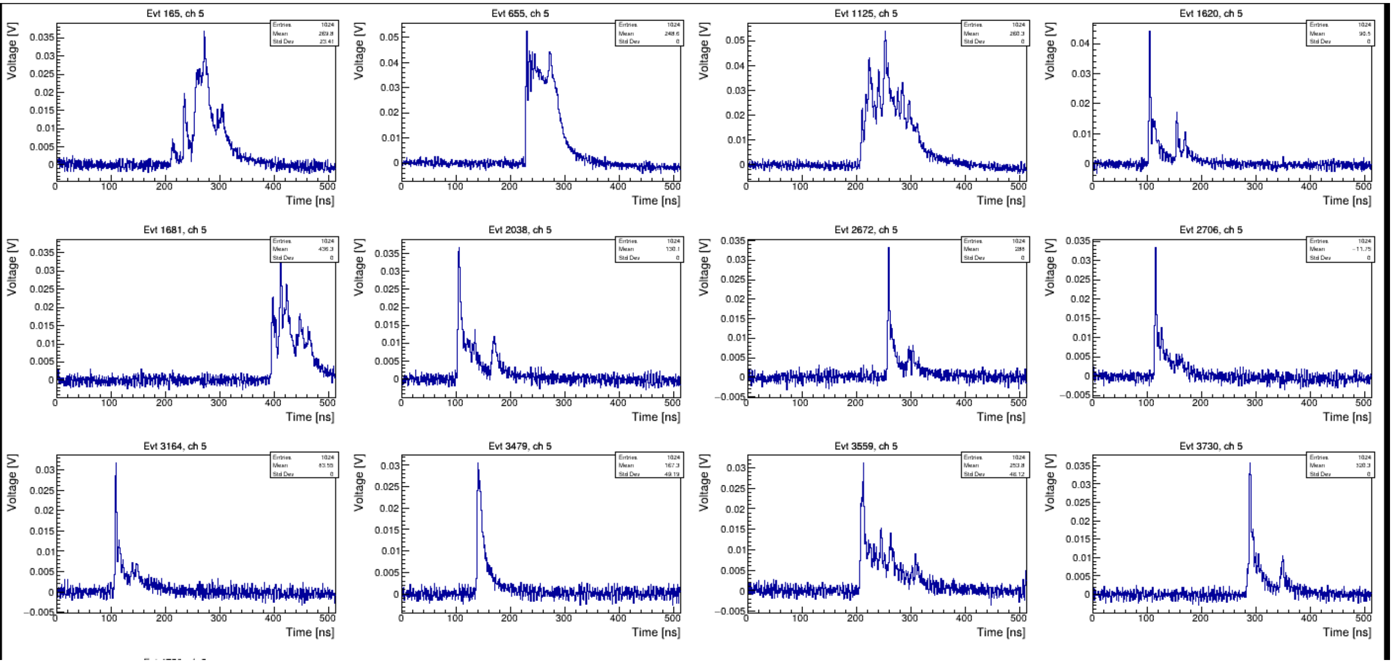




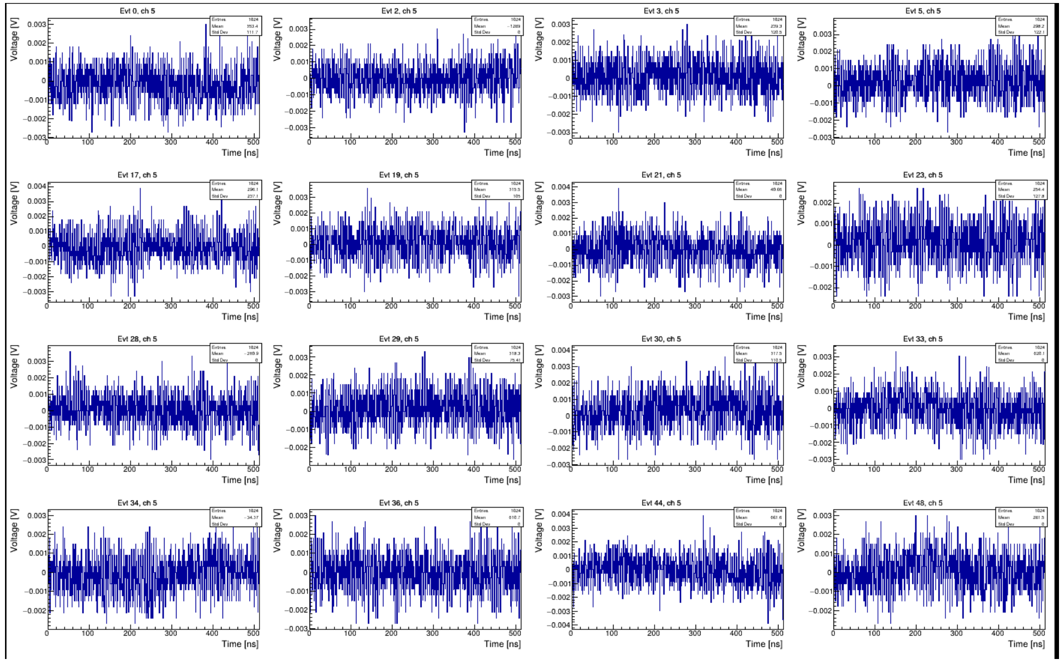
These distributions are related to primary clusters by CNN using threshold (0.9). As a results the mean number of primary clusters and standard deviation less compare to primary clusters (MC)

Beam test data selection for Channel 5

Signal-like data

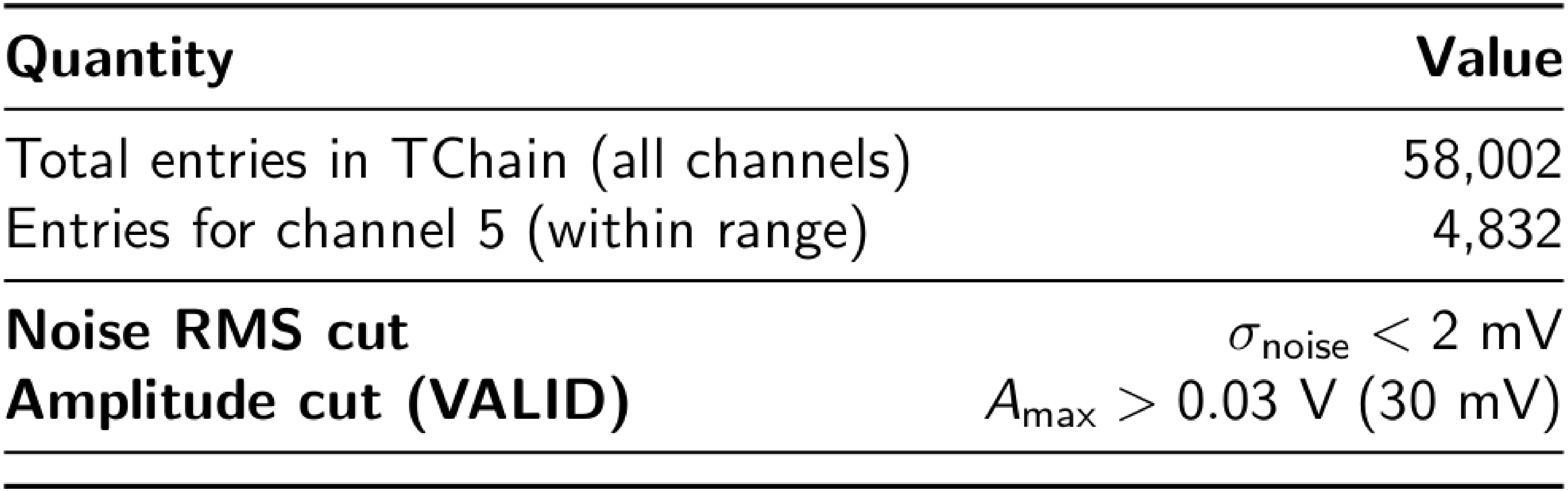


Noise-like data

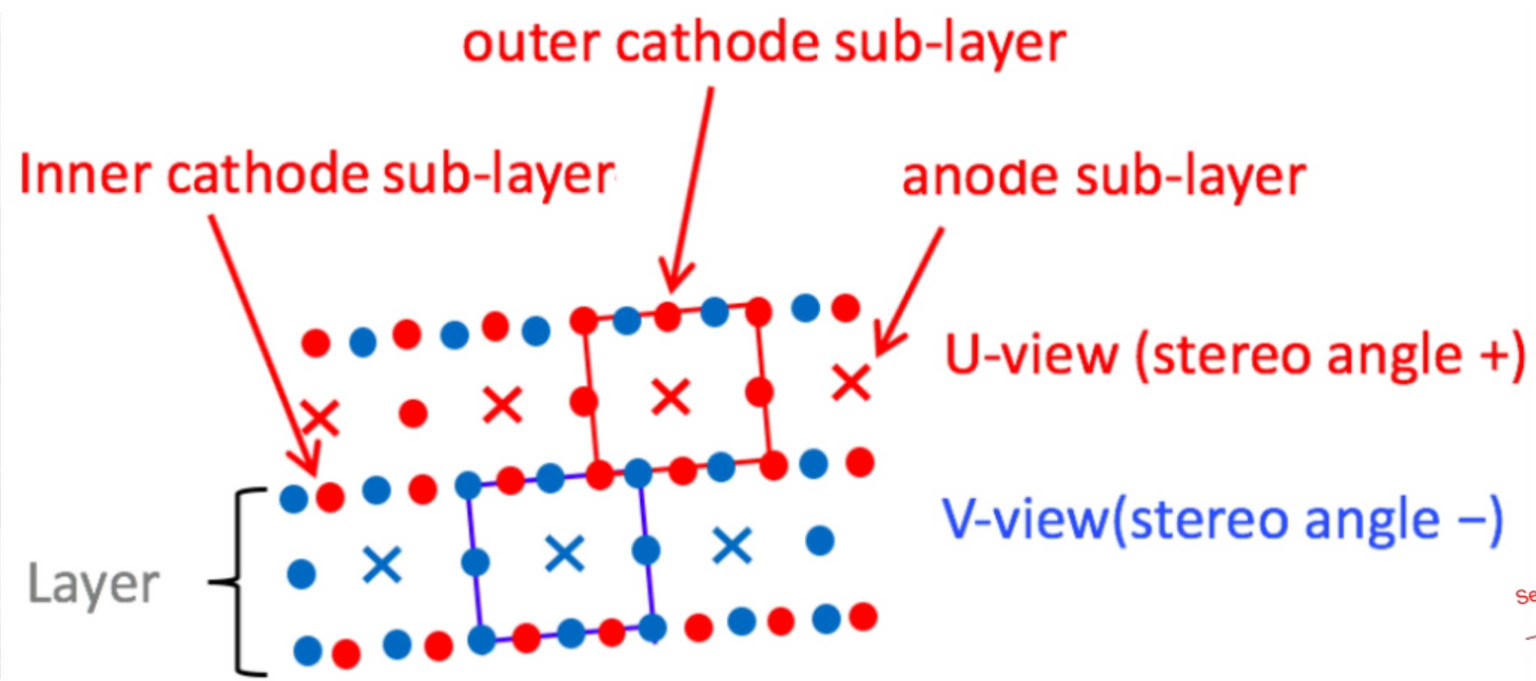


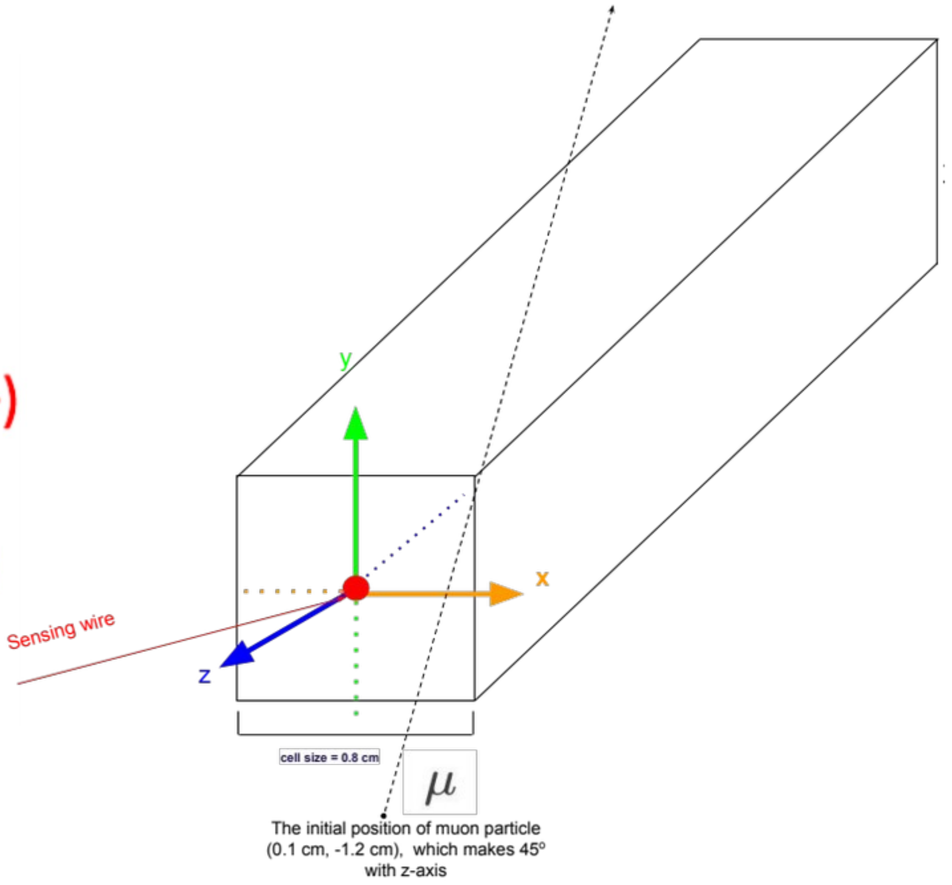
- Signal-like data: peak finding
- Noise-like data: noise information

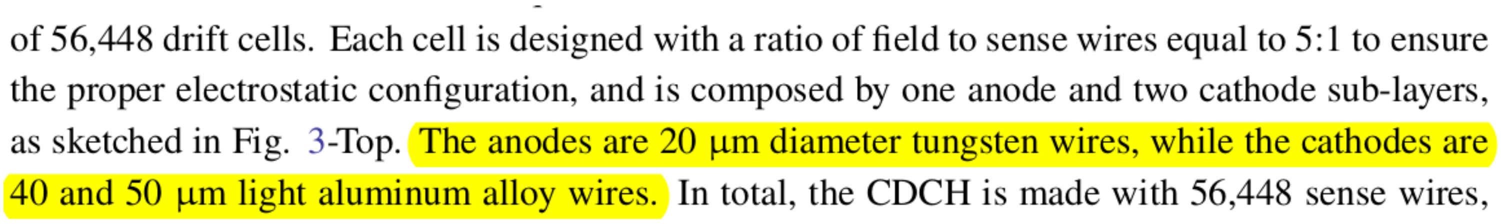
Waveform Selection Logic



Quantity	**Value**
Total entries in TChain (all channels)	58,002
Entries for channel 5 (within range)	4,832
Noise RMS cut	$\sigma_{\text{noise}} < 2$ mV
Amplitude cut (VALID)	$A_{\text{max}} > 0.03$ V (30 mV)



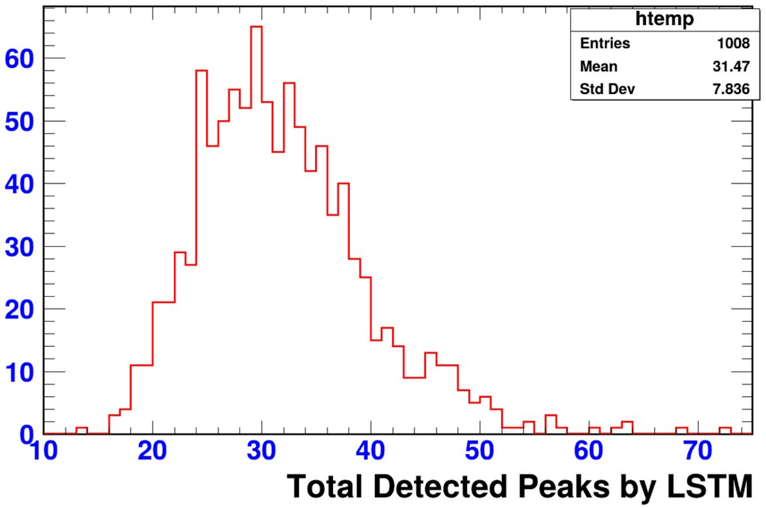

y
x
z
Sensing wire
cell size = 0.8 cm
μ
The initial position of muon particle (0.1 cm, -1.2 cm), which makes 45° with z-axis


of 56,448 drift cells. Each cell is designed with a ratio of field to sense wires equal to 5:1 to ensure the proper electrostatic configuration, and is composed by one anode and two cathode sub-layers, as sketched in Fig. 3-Top. The anodes are 20 μm diameter tungsten wires, while the cathodes are 40 and 50 μm light aluminum alloy wires. In total, the CDCH is made with 56,448 sense wires,

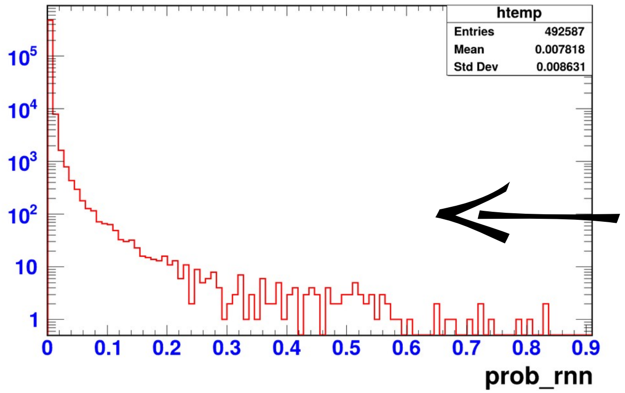
One cell is considered here as one sense wire sorrounded by eight field wires... (Uniform Field)

Applying Untrained NN on beam test data 2022 (180 GeV) Just for Checking

Preliminary Results of Peak Finding distribution

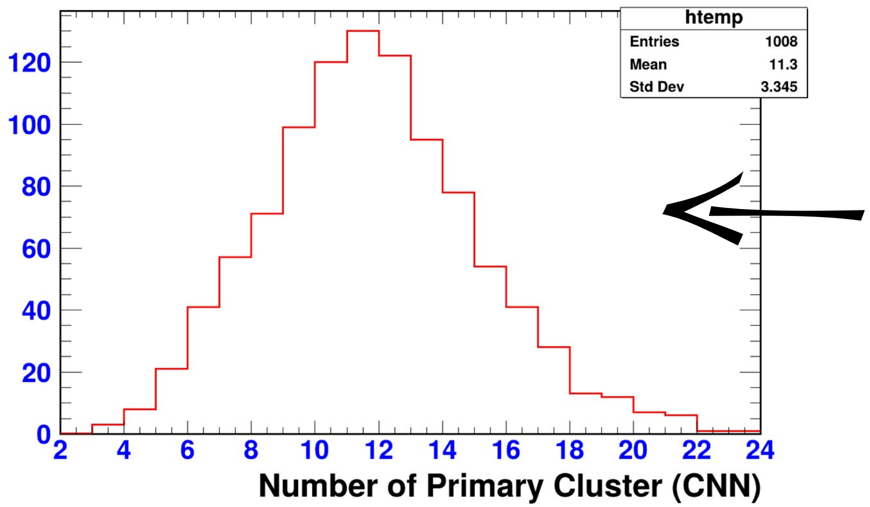


- The above distribution shows us total number of detected peaks (primary and secondary electrons) the

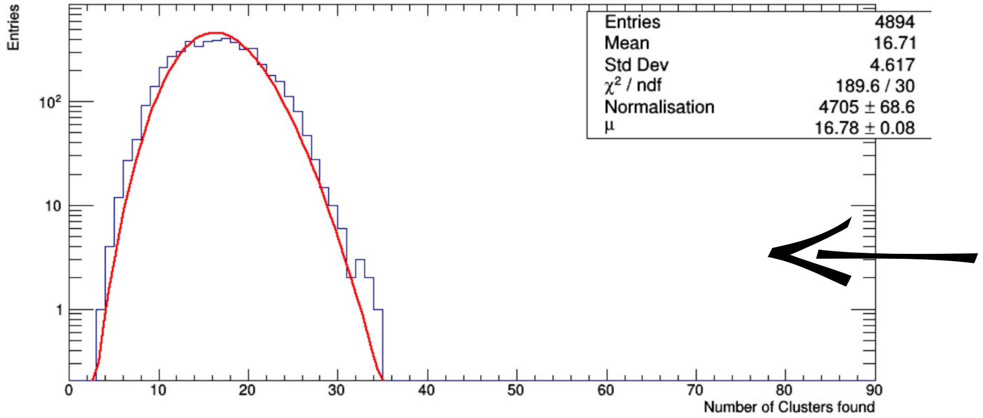


Bad Discrimination of signal and background by using trained LSTM model on untuned simulated waveforms. Because we did not see two main peaks one close to zero represent mostly background events and other close to one represent mostly signal events

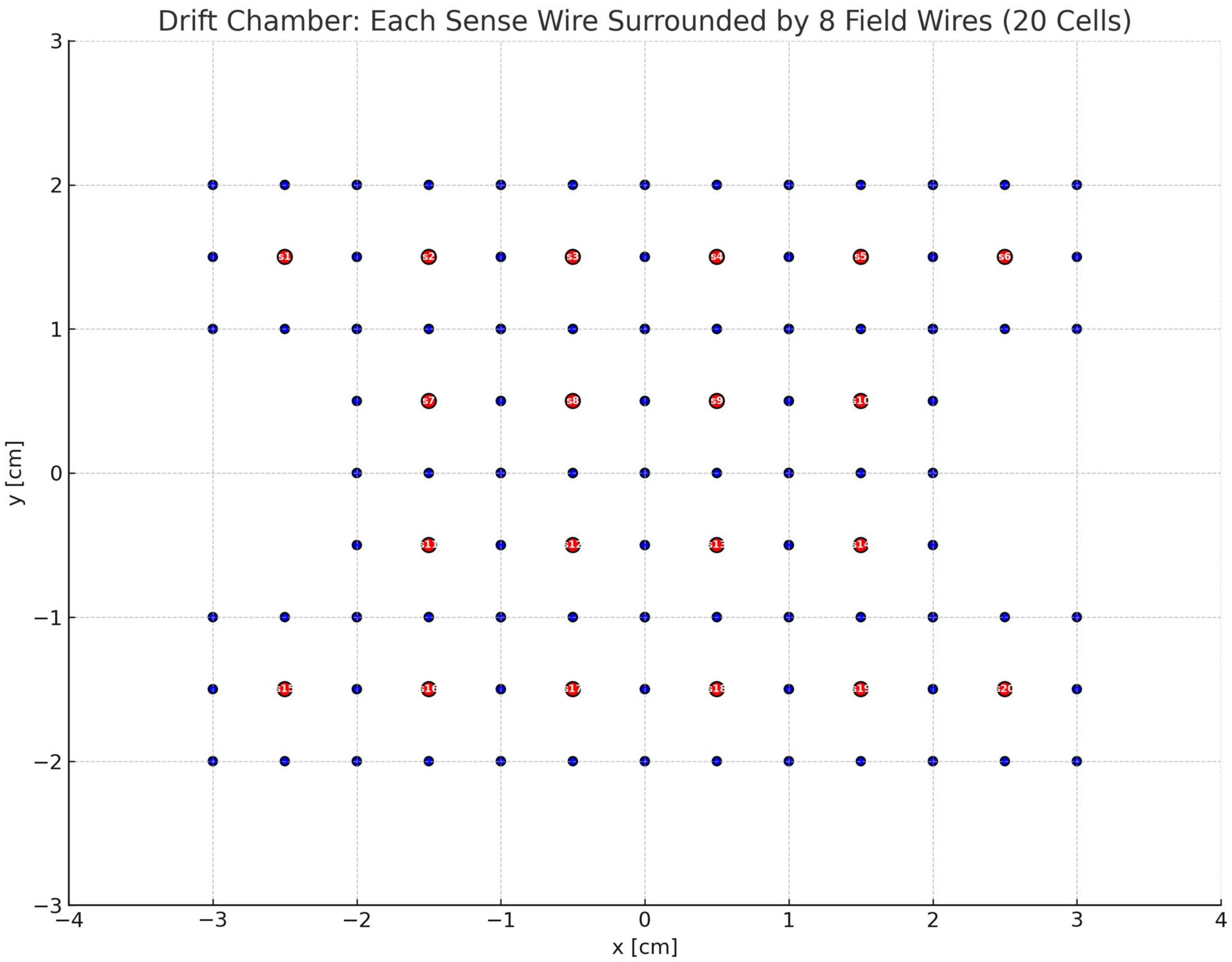
Clusterization



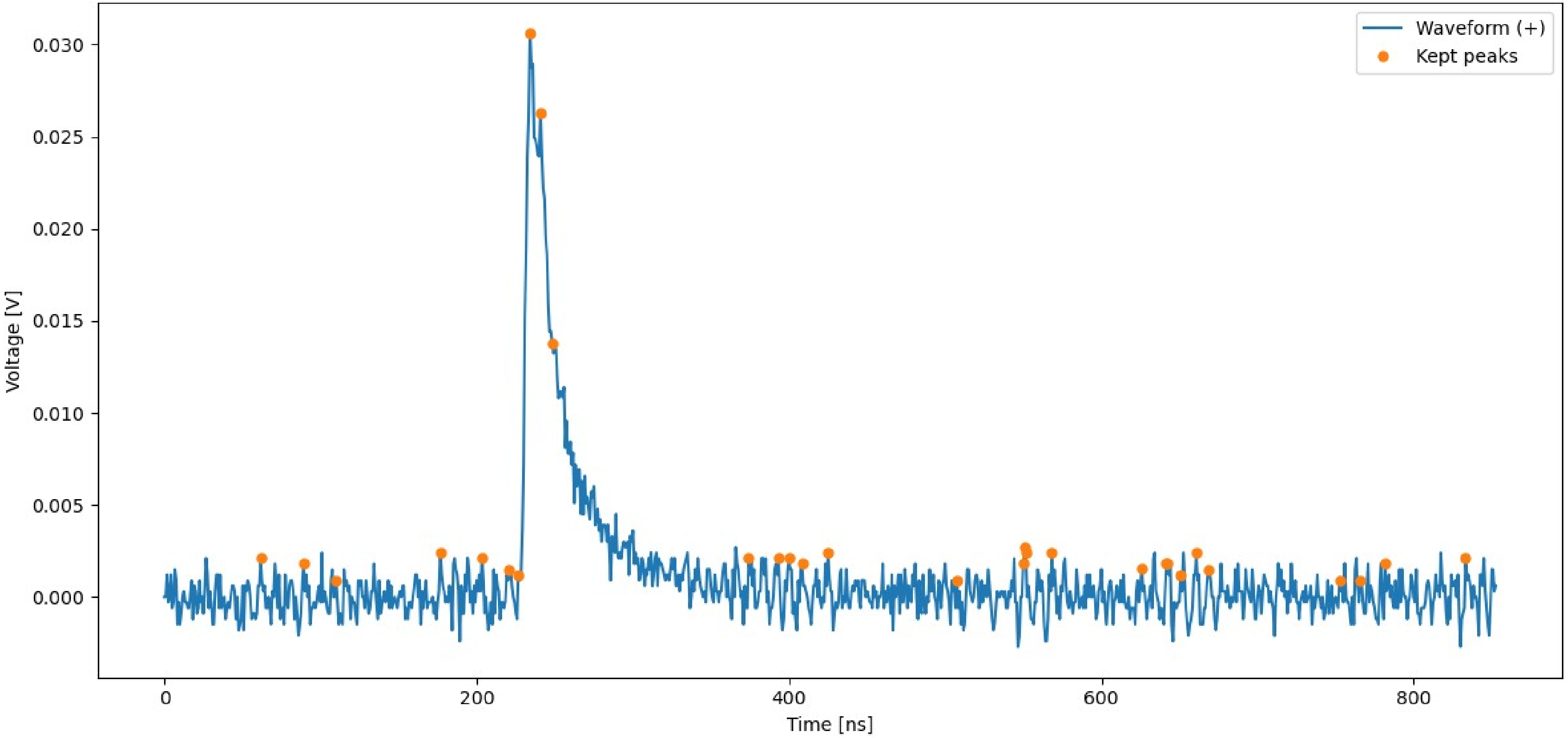
Preliminary CNN Clusterization



RTA Algorithm



Drift Chamber: Each Sense Wire Surrounded by 8 Field Wires (20 Cells)


Waveform (+)
Kept peaks
0.030
0.025
0.020
0.015
0.010
0.005
0.000
Voltage [V]
0
200
400
600
800
Time [ns]

1. Peak Detection Based on Probability Cut:
The detection process involves looping over all events and applying a probability cut to decide whether a peak is considered a valid detection:

Looping Over Events: The script iterates over all entries (events) in the probability file.

Applying the Cut: For each event, it checks if the predicted probability (prob_ml) exceeds the cut threshold (cut), which is set to 0.95/0.65.

Storing Detected Peaks: If the probability exceeds the threshold, the corresponding peak time is stored in the detected_time dictionary, keyed by event number (evtno)

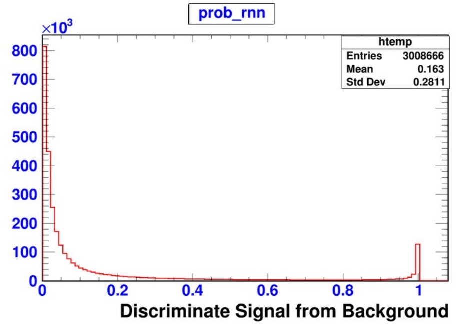
2. Matching Detected Peaks with Truth Data
After detecting peaks, the script matches these peaks with the Monte Carlo (MC) truth data to classify them as primary or secondary:

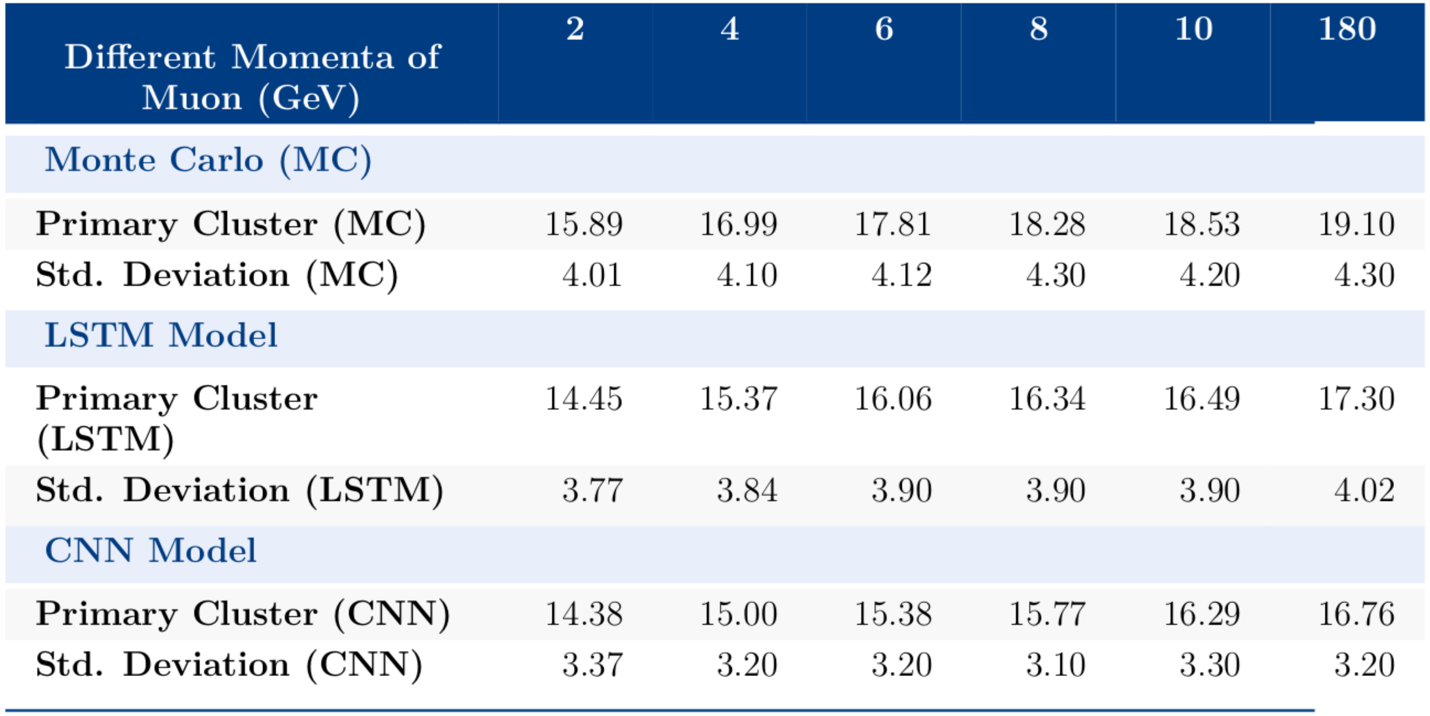
Truth Data: The truth data (truth_time, truth_tag) contains the actual times of primary and secondary peaks, labeled by truth_tag (1 for primary, 2 for secondary).

Matching Function: The match function compares detected peak times with the truth peak times. For each detected peak, it finds the closest truth peak and assigns the corresponding tag (primary or secondary) based on the truth data.

ID Assignment: The id_list array stores the classification of each detected peak as primary (1) or secondary (2)

3. Counting Primary Peaks: After classifying the detected peaks, the script counts how many of them are primary peaks:
Counting: The script iterates over the id_list and increments ncount_pri for every primary peak (tag 1)

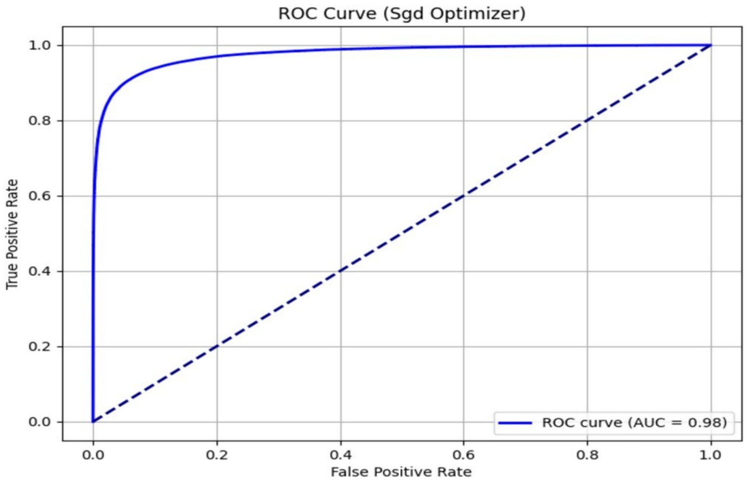


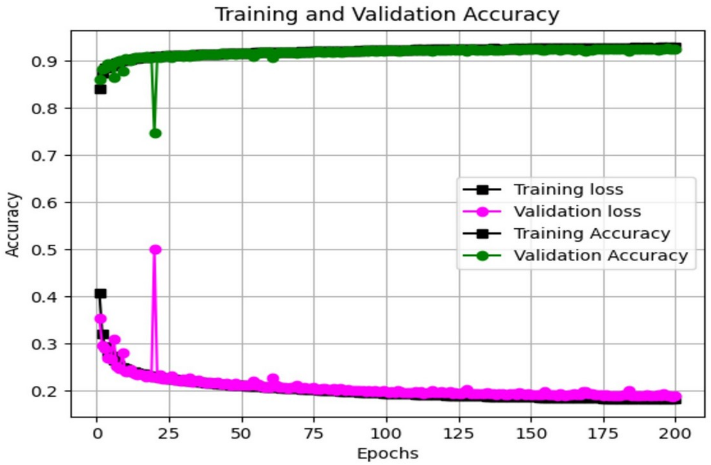


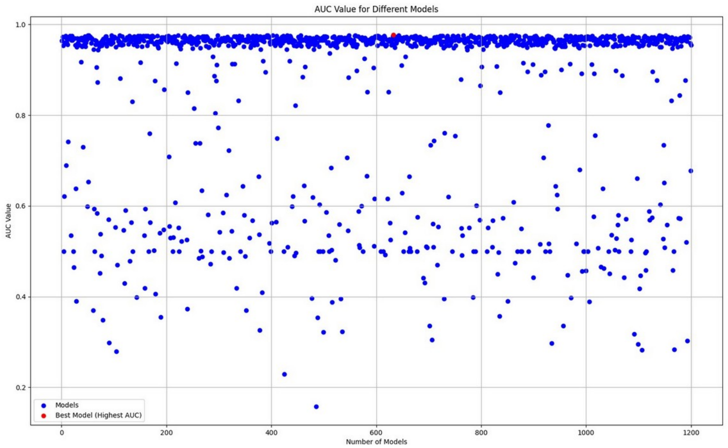
Different Momenta of Muon (GeV)	2	4	6	8	10	180
Monte Carlo (MC)						
Primary Cluster (MC)	15.89	16.99	17.81	18.28	18.53	19.10
Std. Deviation (MC)	4.01	4.10	4.12	4.30	4.20	4.30
LSTM Model						
Primary Cluster (LSTM)	14.45	15.37	16.06	16.34	16.49	17.30
Std. Deviation (LSTM)	3.77	3.84	3.90	3.90	3.90	4.02
CNN Model						
Primary Cluster (CNN)	14.38	15.00	15.38	15.77	16.29	16.76
Std. Deviation (CNN)	3.37	3.20	3.20	3.10	3.30	3.20

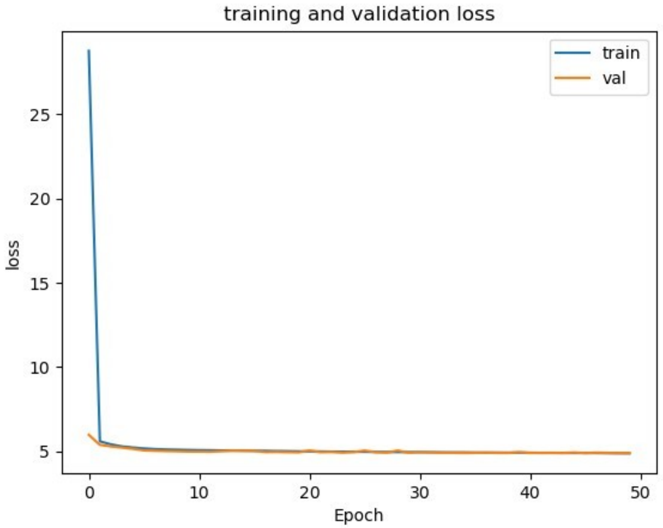
Table 1: Primary cluster means and standard deviations from MC, LSTM, and CNN across different muon momenta.

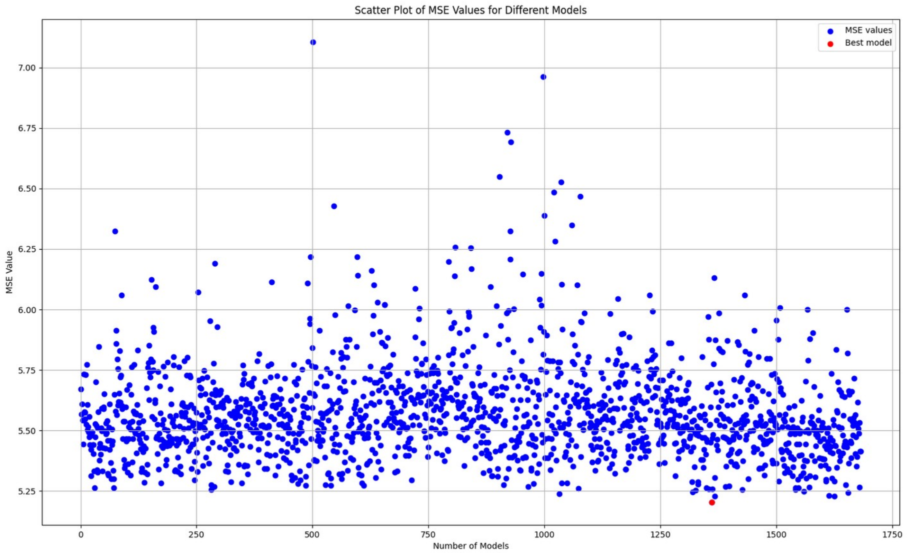
Best LSTM Peak Finding (Above 1st row) and CNN Regression Model (Below 2nd Row)



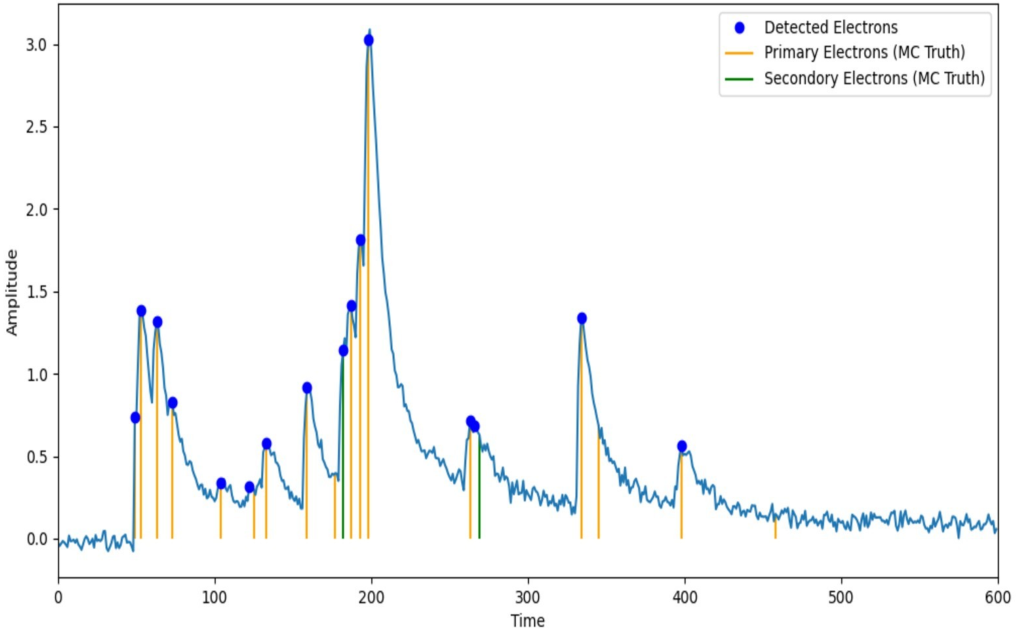


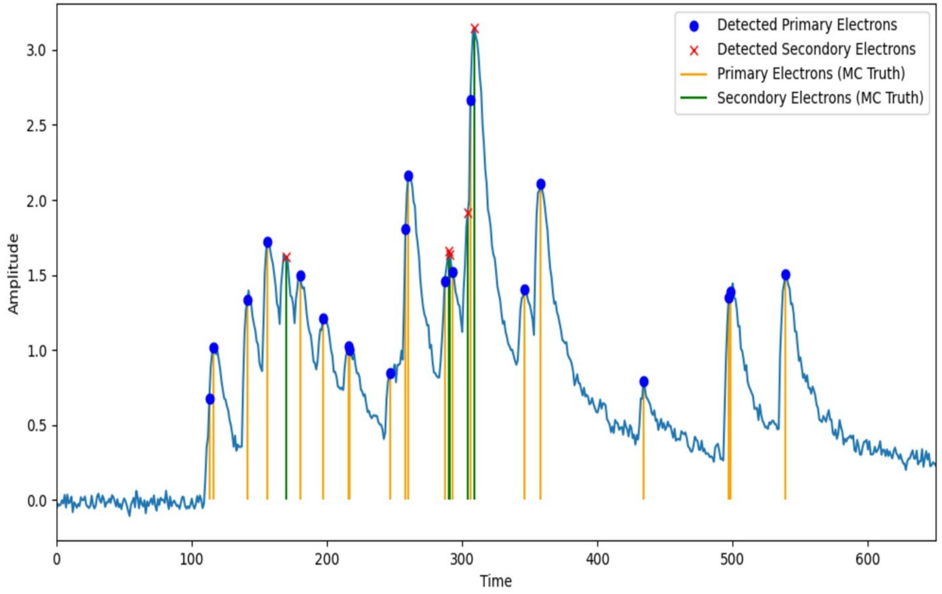


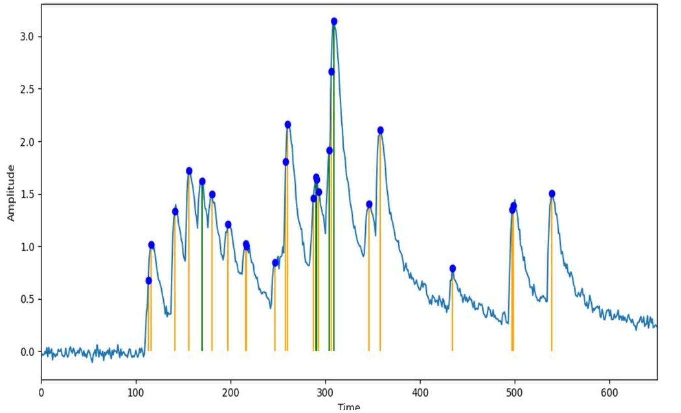


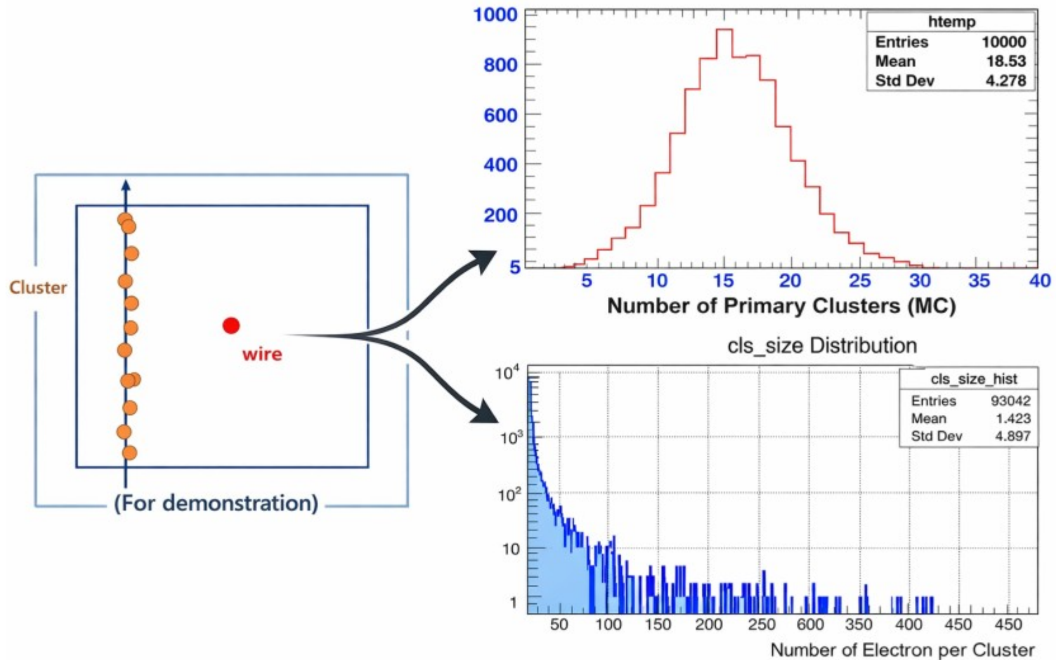


Simulated Peak Finding Waveforms









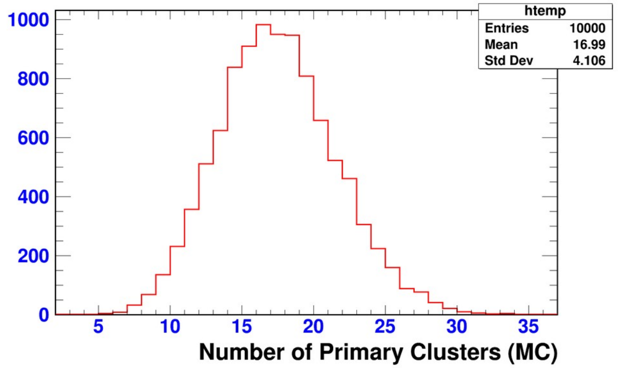
$$N_{\text{primary clusters}} \sim \text{Poisson}(\bar{N}_p)$$

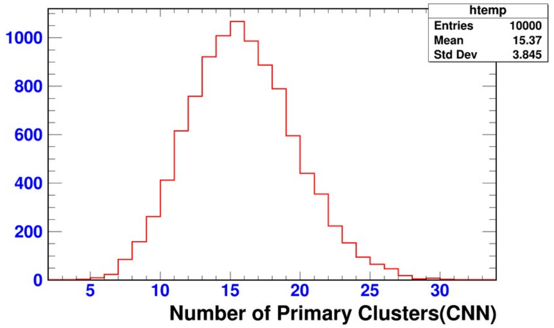
with the mean

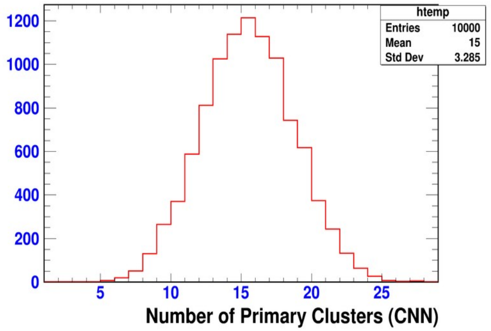
$$\bar{N}_p = \lambda L$$

- λ: average primary-cluster density (clusters per cm)
- L: track length in the gas (cm)

Final results of the reconstruction for 4 GeV on the base of different selection cuts







Different selection cuts for 4 GeV	MC	σ of MC	Primary Cluster(LSTM)	σ of (LSTM)	CNN	σ of (CNN)
0.55	16.99	4.1	15.37	3.84	15.	3..2
0.65	16.99	4.1	15.05	3.78	14.9	3.3
0.85	16.99	4.1	14.05	3.5	13.77	3.2
0.95	16.99	4.1	12.74	3.32	12.19	2.99