Particle Reconstruction in Range System

V. Zel on behalf of SPD Muon Group

Outline

- Data preprocessing
- Clustering
  - Metrics
  - DBSCAN
- Particle identification (classification)
  - Metrics
  - Features
  - Decision tree
  - Random forest
  - XGBoost
  - Convolution neural network
- Conclusions

Particle reconstruction in Range System

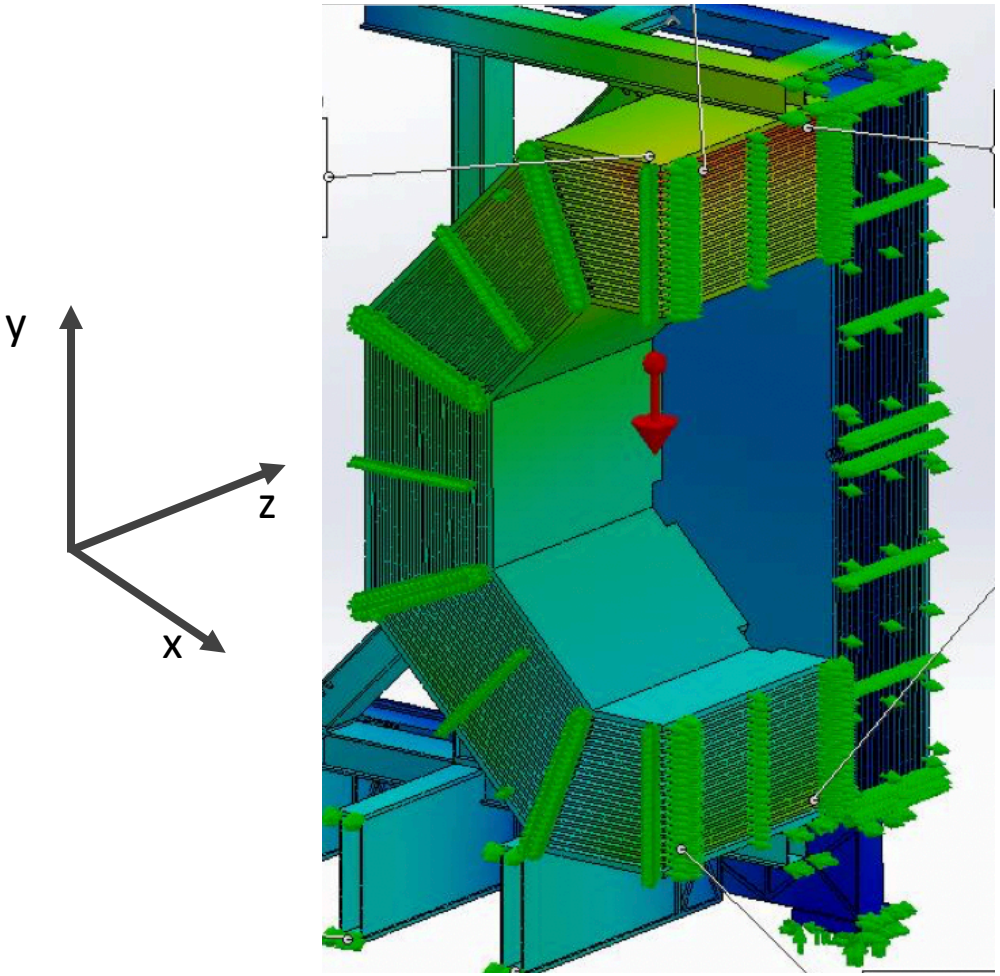
Information available from Range System:

- hits in Barrel: *(x, y)* of wires at layers and *z* of strips
- hits in EndCaps: *(y, z)* of wires and *x* of strips

Two steps of particle reconstruction:

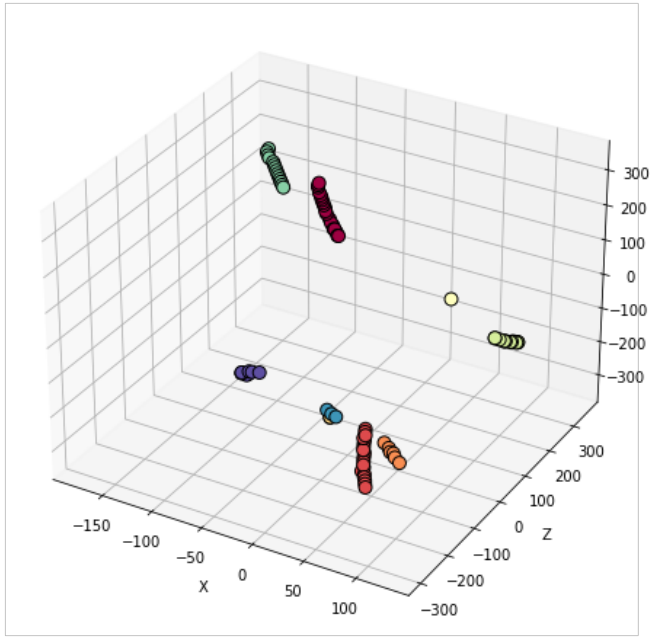
1. Clustering - forms group of hits (clusters)
2. Particle identification (cluster labeling)

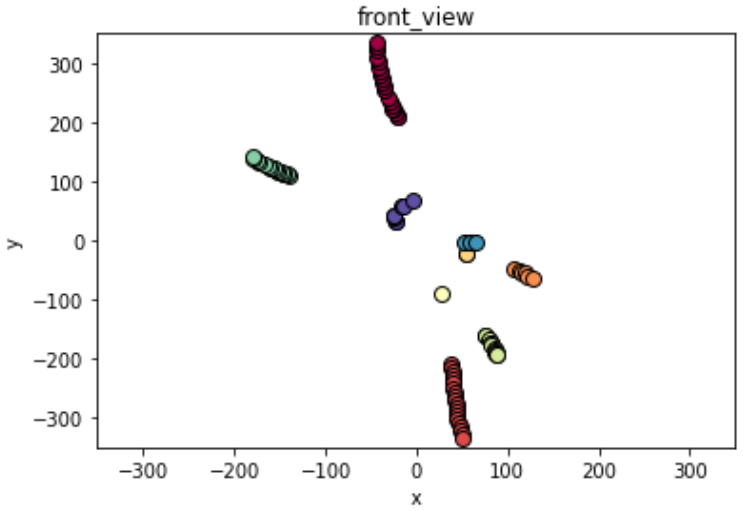
Work is based on the use 50k $J/\psi \to \mu\mu$ Monte Carlo events proton-proton collisions at a beam energy of E = 27 GeV

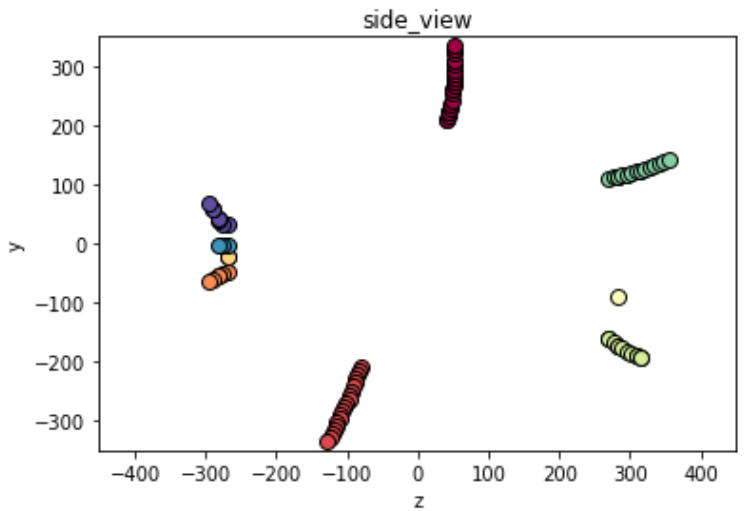


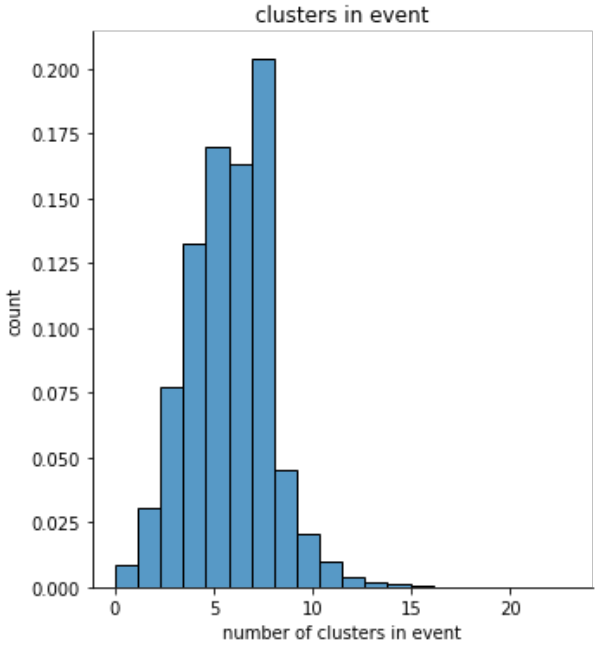
Cross section of the SPD RS

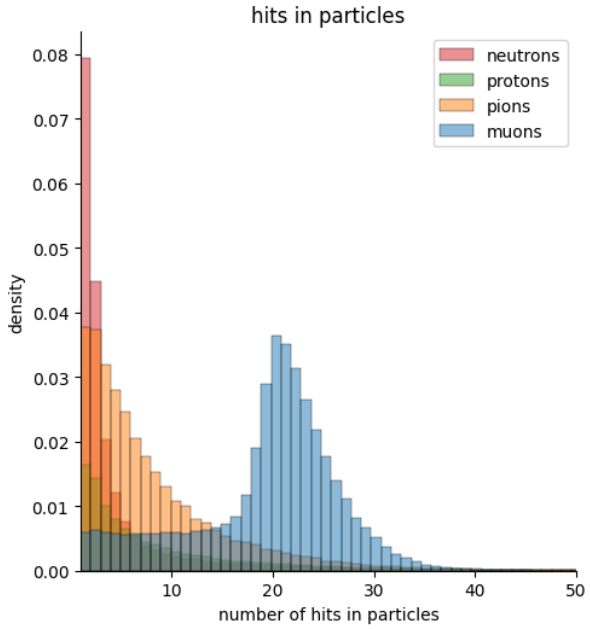
Data preprocessing

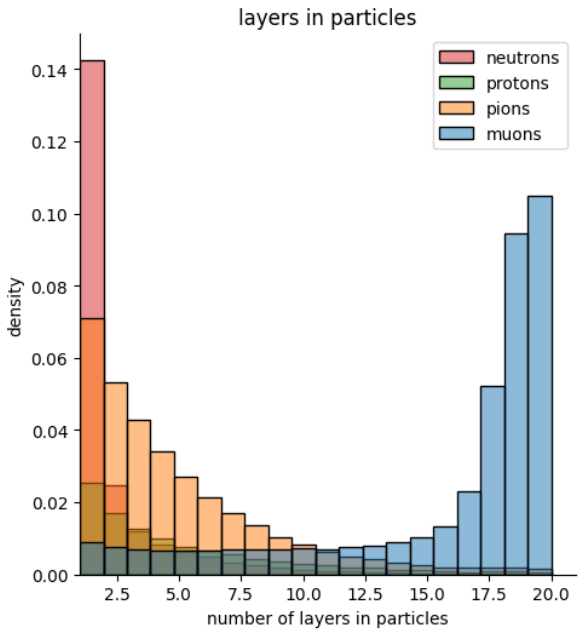












Clustering

Clustering is unsupervised machine learning technique that groups data points into clusters based on their similarities.

Performance metrics:

$$Purity = \frac{\sum_i N_{i,hits}^{correct}}{N_{hits}^{total}}$$

$$V\text{-}measure = \frac{(1+\beta)*homogeneity*completeness}{(\beta*homogeneity+completeness)},$$

where by default $\beta = 1$.

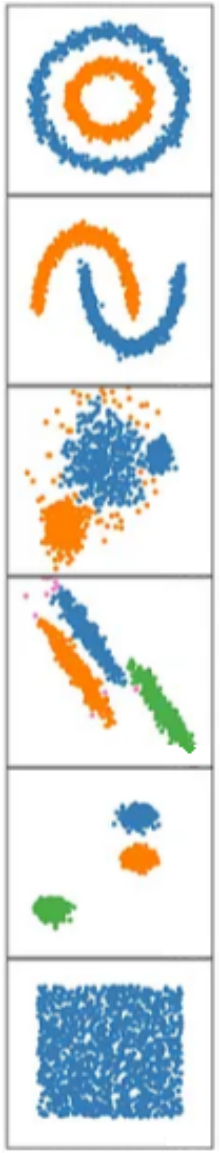
- *homogeneity*: each cluster contains only members of a single class
- *completeness*: all members of a given class are assigned to the same cluster

DBSCAN *(Density Based Spatial Clustering of Application with Noise)*:

- Can identify clusters of arbitrary shapes and sizes;
- It does not require a pre-set number of clusters;
- Handle noise and outliers in data.

Input parameters :
ε - distance within which two points can be considered to belong to the same cluster;

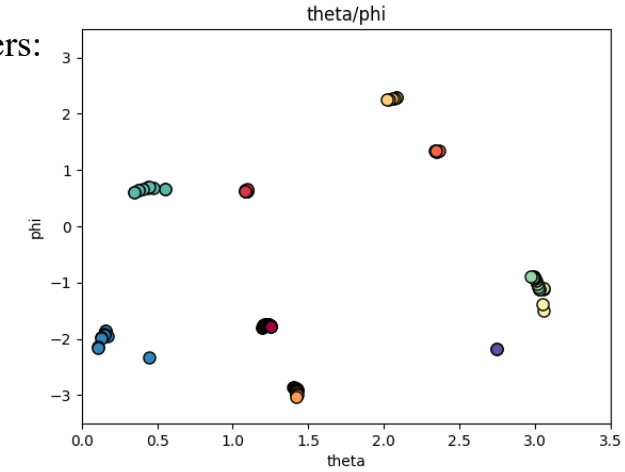
MinPts - minimum number of points to define a cluster.

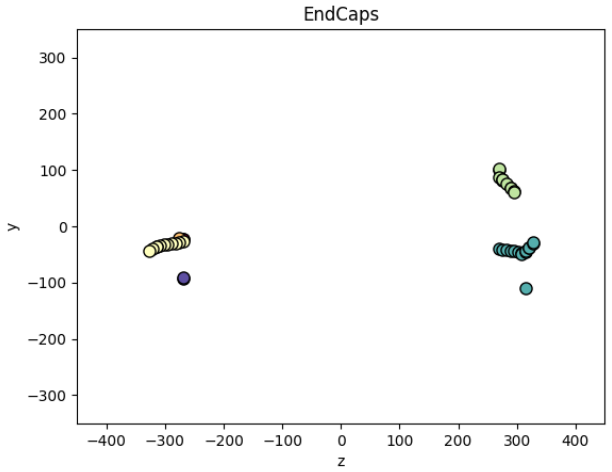


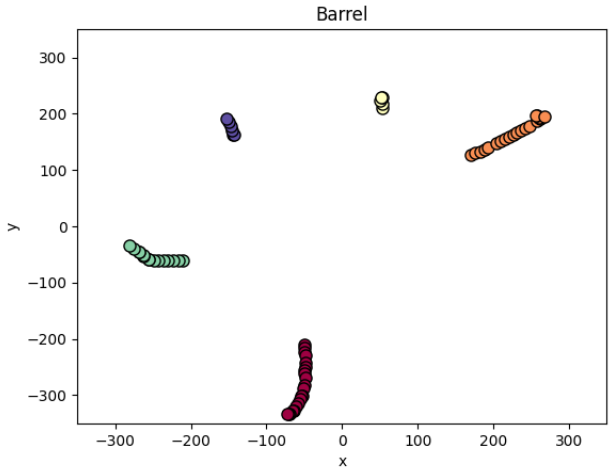
DBSCAN perfomance

DBSCAN result for single event:

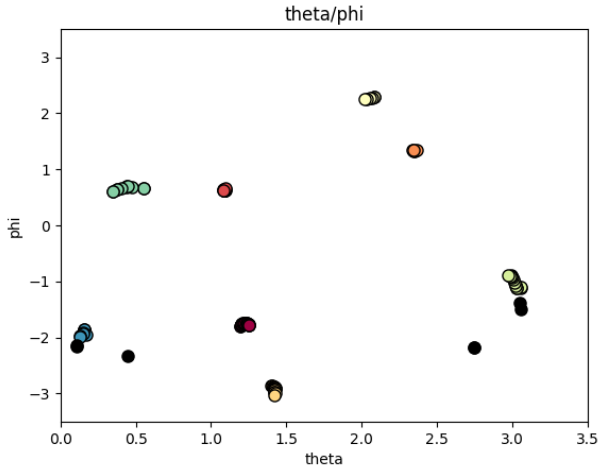
Real clusters:

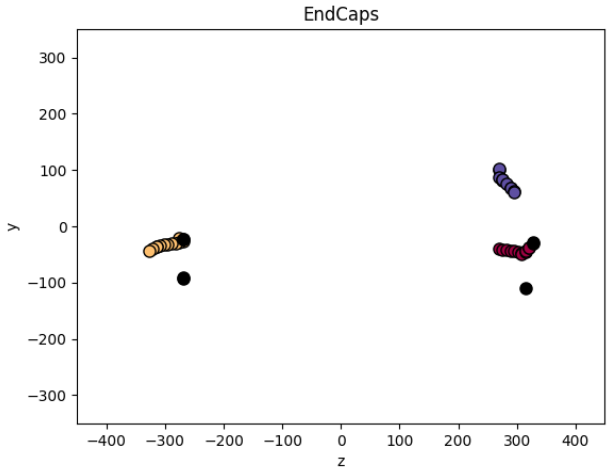


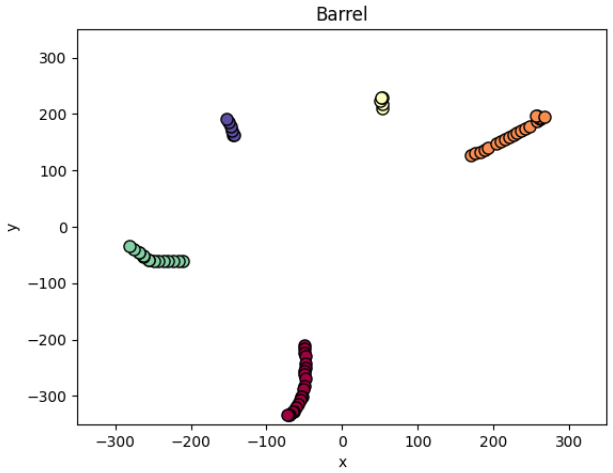




DBSCAN clusters:







DBSCAN metrics:

Homogeneity: 0.98
Completeness: 0.98
V-measure: 0.98
Purity: 0.97

Particle identification

Classification is a common task in machine learning that involved predicting the class or category of a given input data point

Performance metrics:

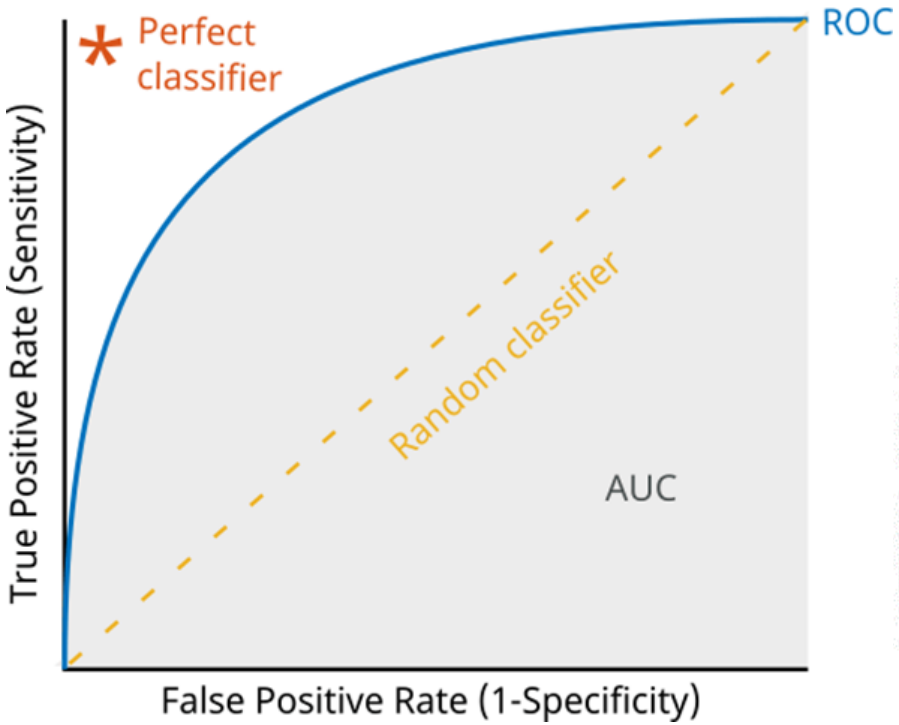
$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$f1 = \frac{2*precision*recall}{precision+recall}, \text{ where:}$$

$$precision = \frac{TP}{TP+FP},$$

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

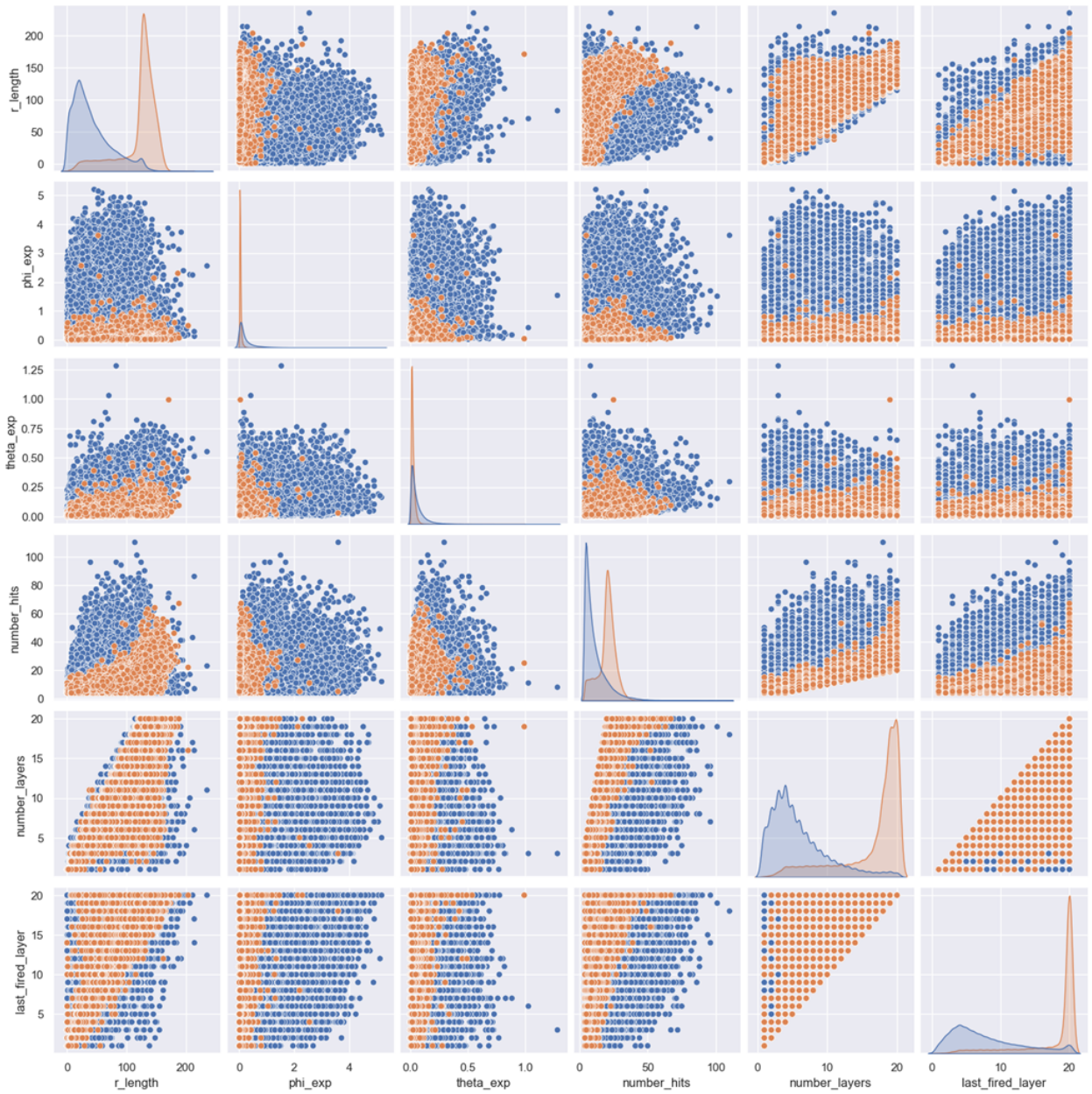
AUC-ROC

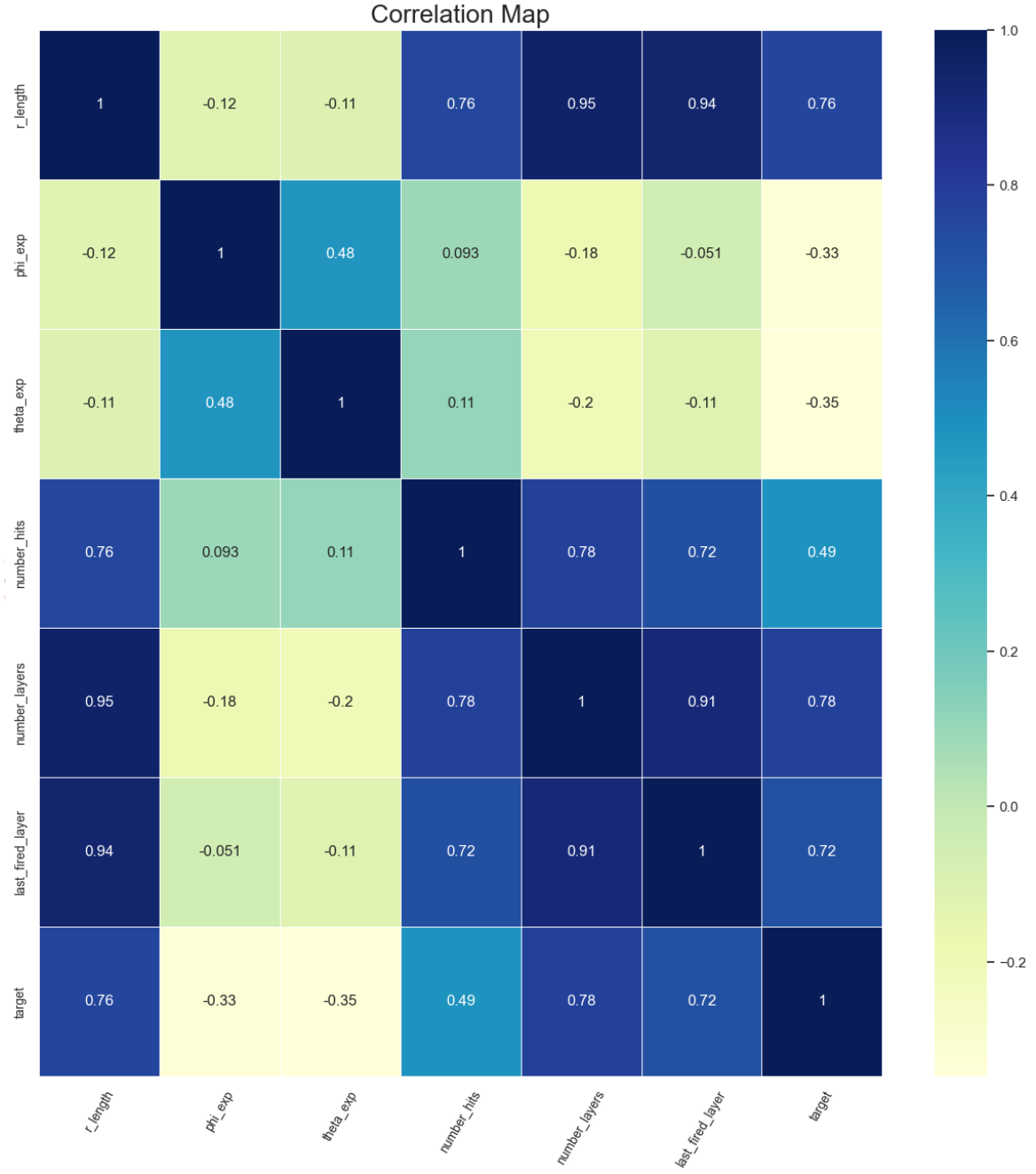


ACTUAL VALUES	PREDICTIVE VALUES: POSITIVE (1)	PREDICTIVE VALUES: NEGATIVE (0)
POSITIVE (1)	TP	FN
NEGATIVE (0)	FP	TN

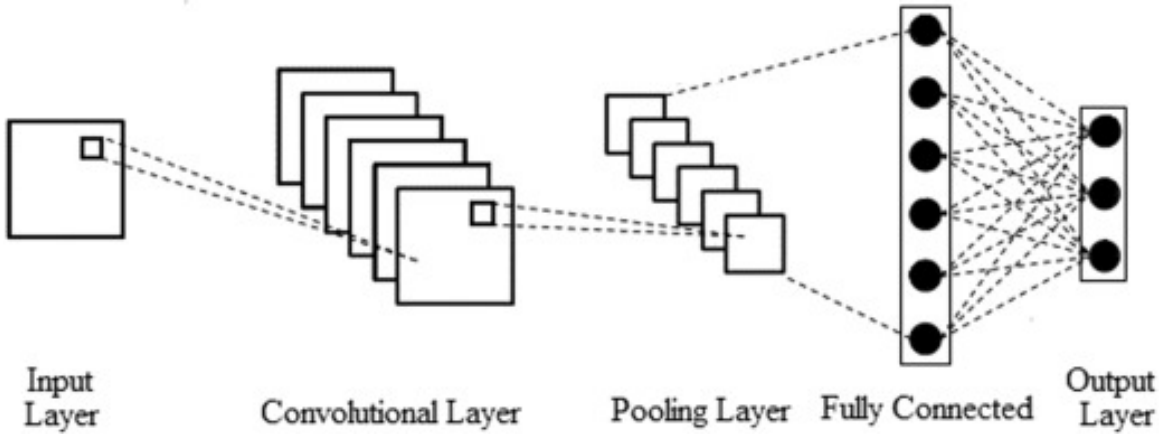
Algorithms used: ***Decision Tree, Random Forest, XGBoost, CNN.***

Features


r_length
phi_exp
theta_exp
number_hits
number_layers
last_fired_layer


Correlation Map
r_length 1 -0.12 -0.11 0.76 0.95 0.94 0.76
phi_exp -0.12 1 0.48 0.093 -0.18 -0.051 -0.33
theta_exp -0.11 0.48 1 0.11 -0.2 -0.11 -0.35
number_hits 0.76 0.093 0.11 1 0.78 0.72 0.49
number_layers 0.95 -0.18 -0.2 0.78 1 0.91 0.78
last_fired_layer 0.94 -0.051 -0.11 0.72 0.91 1 0.72
target 0.76 -0.33 -0.35 0.49 0.78 0.72 1
r_length
phi_exp
theta_exp
number_hits
number_layers
last_fired_layer
target

Convolutional Neural Network



Convolutional Neural Networks (CNNs) are a type of deep learning algorithm commonly used for image and video recognition tasks.

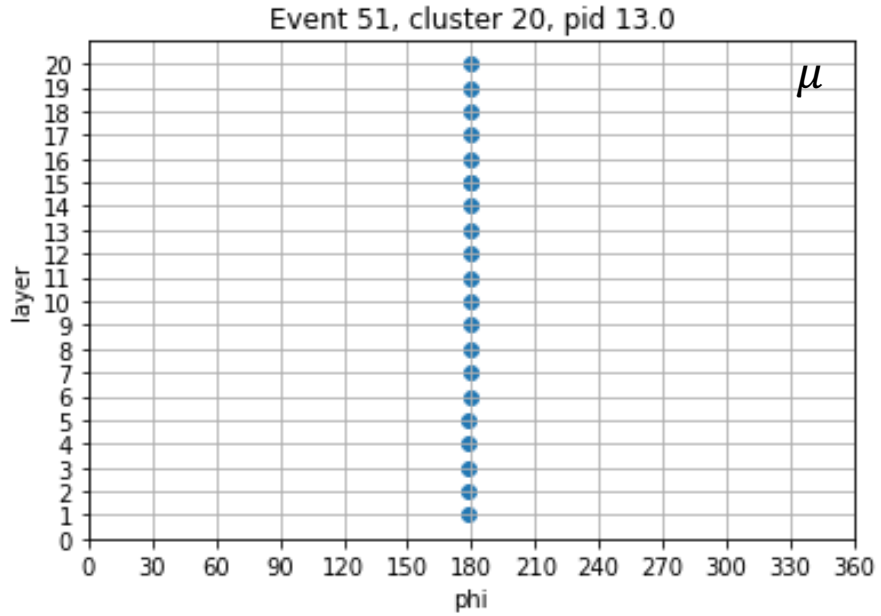
Advantages:

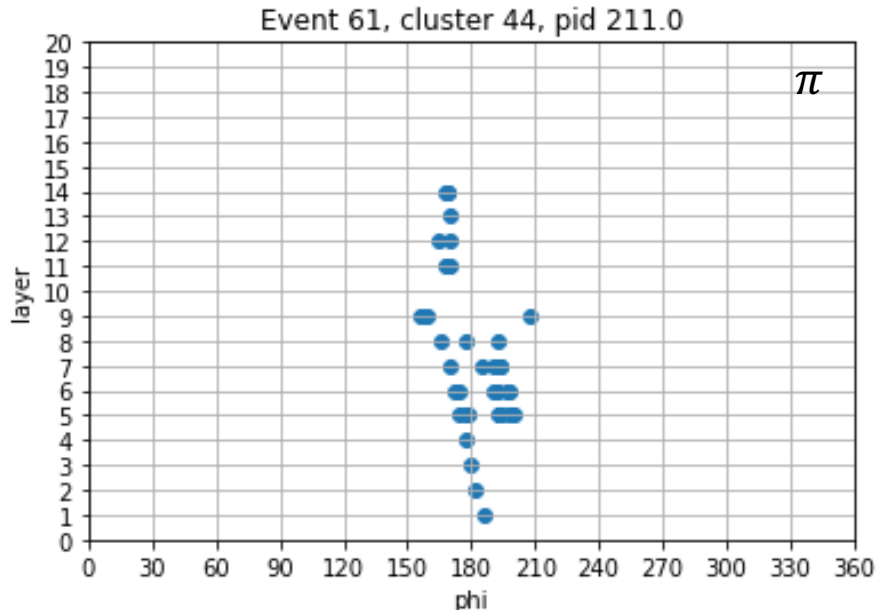
- ability to capture complex patterns and relationships
- robustness to variations in input data.

Disadvantages:

- large amount of training data
- longer training time
- difficulty in interpreting the learned features.

Example of 2d image (20x360) input





Algorithms performance

	Decision Tree	Random Forest	XGBoost	CNN
Precision	0.94	**0.95**	0.94	0.89
Recall	0.90	0.89	0.90	**0.96**
Accuracy	0.92	0.92	0.92	0.92
F1-score	0.92	0.92	0.92	0.92
AUC-ROC	0.97	0.97	0.98	-

Conclusions

1. Application of the machine learning methods for muon/hadron separation has shown the promising results.
2. The performance of DBSCAN algorithm in the clustering analysis has been evaluated. Using the optimal parameters we obtained purity of 0.97 and v-measure of 0.98.
3. Decision tree, Random Forest, XGBoost and convolution neural network were tested as classifiers. In general, first three algorithms have shown the similar results (precision ~ 0.94-0.95, recall ~ 0.89-0.90). CNN have shown a good result in recall metric – 0.96. However, there is a potential of improving the quality of classification using Random Forest and CNN methods.

Backup

Contribution to the error of the Random Forest algorithm:

	Muon versus Pion	**Muon versus Proton**	**Muon versus Rest (pid: 130, 311, 321…)**	**Muon versus Neutron**
Precision	0.93	0.94	**0.98**	**0.99**
Recall	0.89	0.89	0.89	0.89
Accuracy	0.91	0.91	0.93	0.94
F1-score	0.91	0.91	0.93	0.94
AUC-ROC	0.97	0.97	0.99	0.99
Percentage of cases	55%	16.4%	13.6%	15%