Particle reconstruction in Range System

V.V. Zel
on behalf of SPD Muon Group

Outline

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

Particle reconstruction in Range System

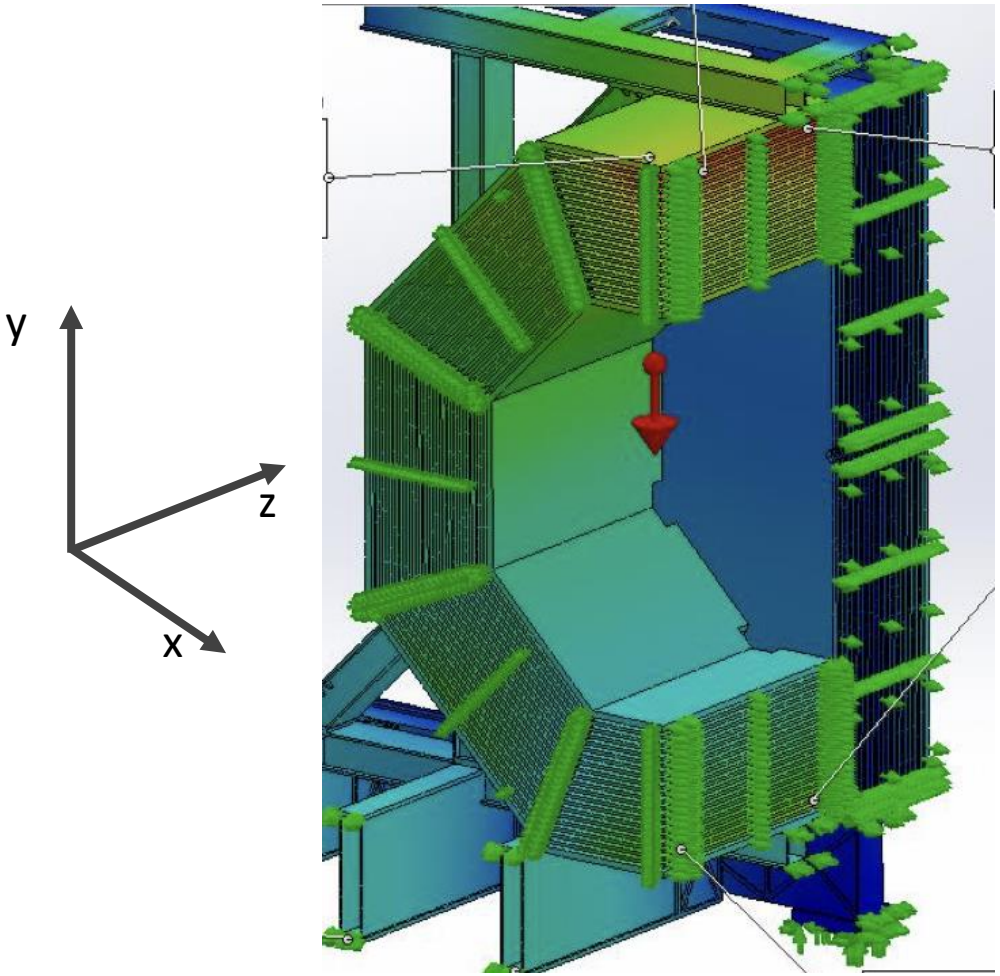
Information available from Range System:

- 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 30k $J/\psi \to \mu\mu$ Monte Carlo events

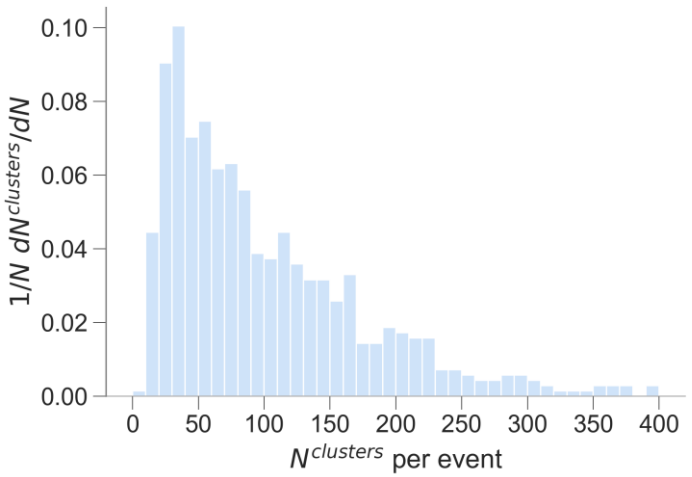


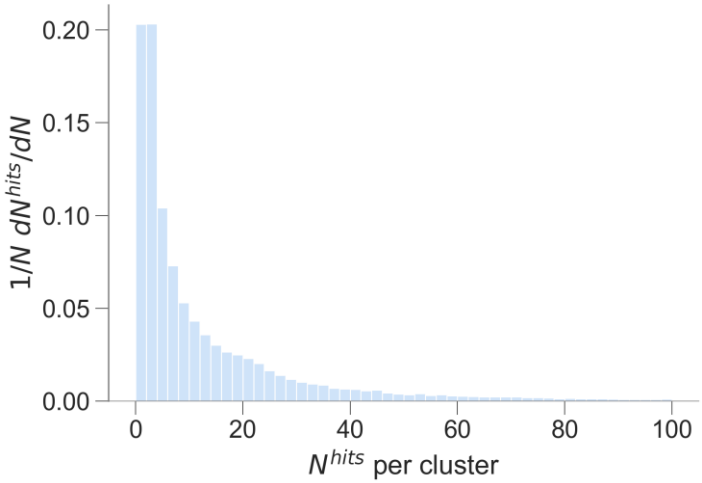
Cross section of the SPD RS

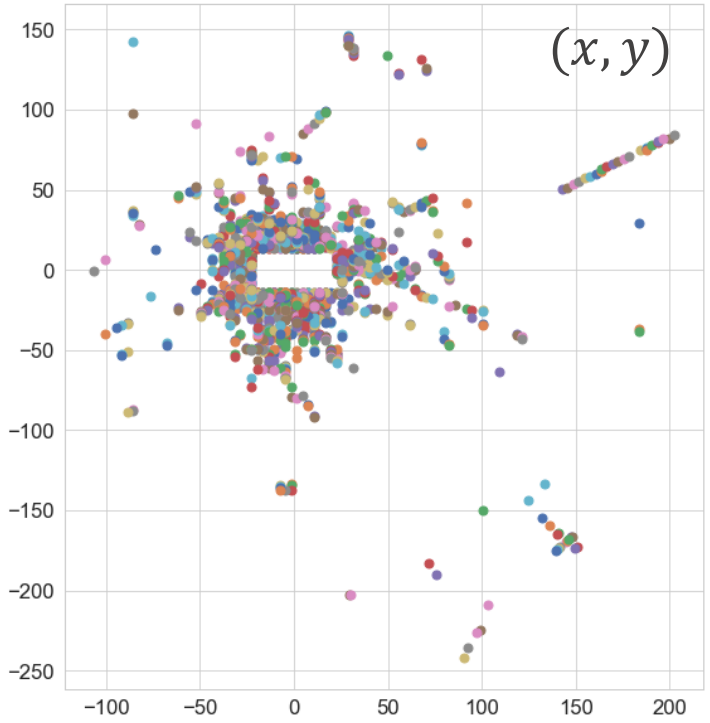
Data preprocessing

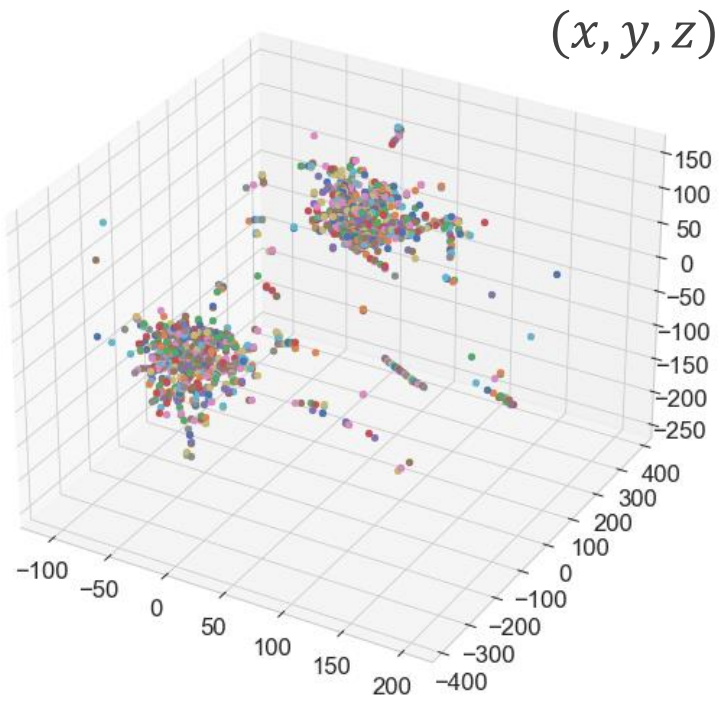
1) High hit density (mostly in forward region) =>
Application of the threshold on $\theta(\eta)$ is needed to remove the regions with high hit density in the vicinity of the beam

2) ~40% clusters have ≤4 hits =>
Clusters can be labeled as noise









Clustering

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

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
- Two input parameters: ε (distance within which two points can be considered to belong to the same cluster) and MinPts (minimum number of points to define a cluster)

Performance metrics:

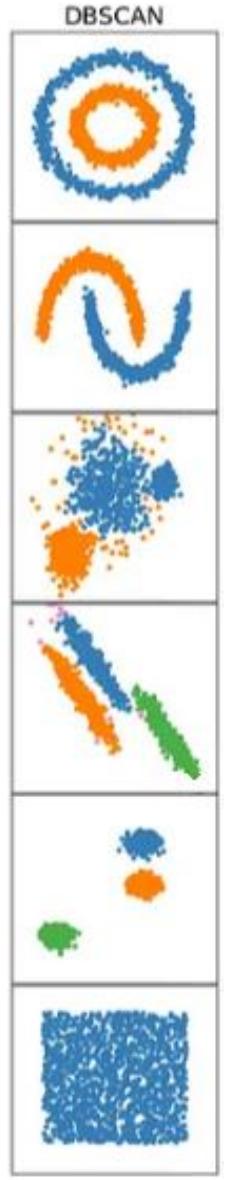
Purity: $P = \frac{\sum_i N_{i,hits}^{correct}}{N_{hits}^{total}}$

V-measure - harmonic mean between the homogeneity and completeness, where:

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

$$v = \frac{(1+\beta)*homogeneity*completeness}{(\beta*homogeneity+completeness)},$$

where by default $\beta = 1$.

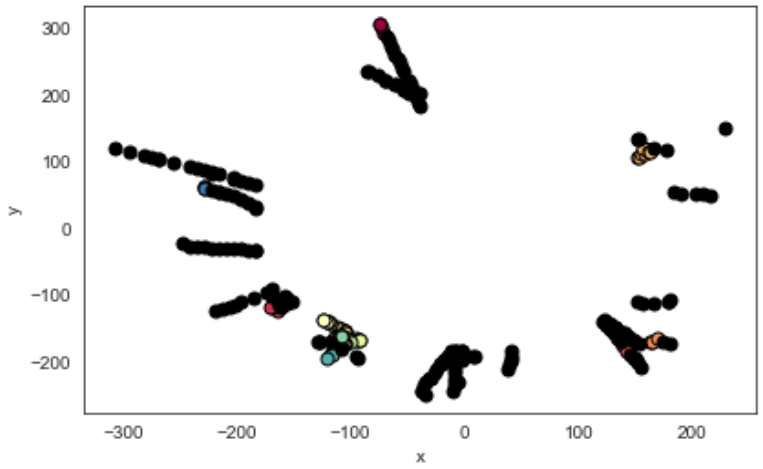


Cluster forms

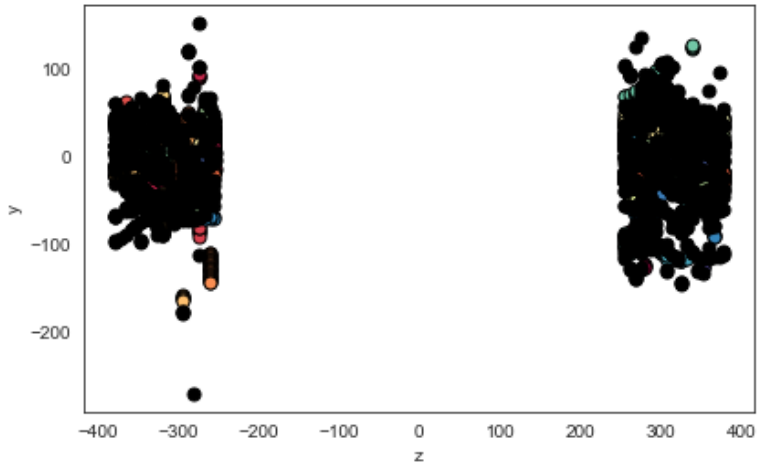
DBSCAN result for single event

all

Barrel

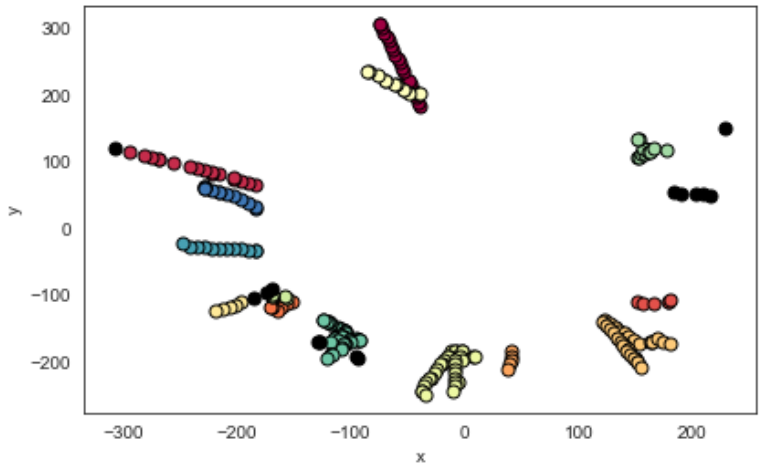


EndCaps

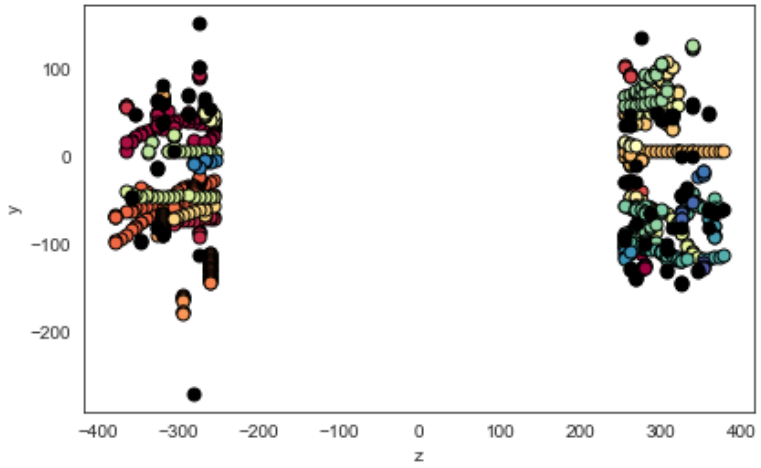


$\pi/16 < \theta < 15\pi/16$
$(|\eta| < 2.4)$

Barrel

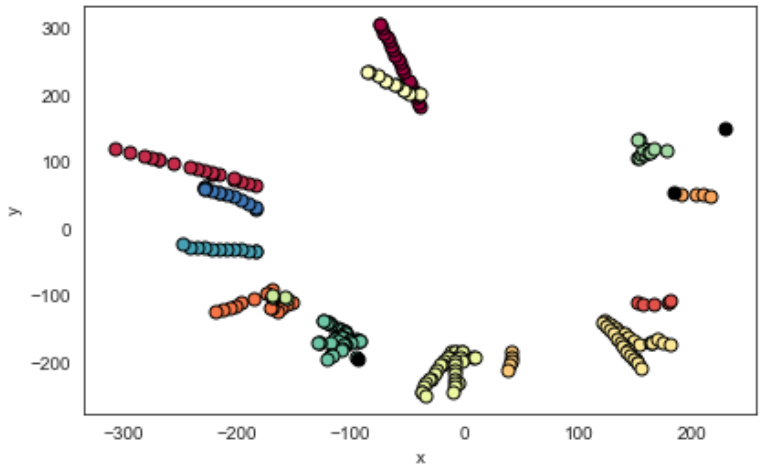


EndCaps

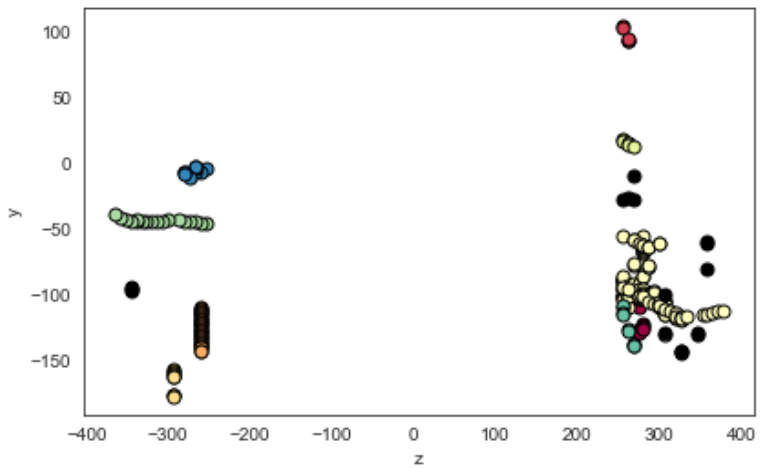


$\pi/8 < \theta < 7\pi/8$
$(|\eta| < 1.6)$

Barrel



EndCaps



DBSCAN performance

	threshold	purity	V-measure	muon loss(%)	Nclust relative err(%)
DBSCAN (θ,φ)	—	0.58	0.71	0	0.37
DBSCAN (θ,φ)	$\pi/16 < \theta < 15\pi/16$ ($\lvert\eta\rvert < 2.4$)	0.89	0.93	~30	0.13
DBSCAN (θ,φ)	$\pi/8 < \theta < 7\pi/8$ ($\lvert\eta\rvert < 1.6$)	0.96	0.97	~60	0.08
DBSCAN (x,y,z)	—	0.67	0.75	0	0.18
DBSCAN (x,y,z)	$\boldsymbol{\pi/16 < \theta < 15\pi/16}$ $\boldsymbol{(\lvert\eta\rvert < 2.4)}$	**0.92**	**0.94**	**~30**	**0.15**
DBSCAN (x,y,z)	$\pi/8 < \theta < 7\pi/8$ ($\lvert\eta\rvert < 1.6$)	0.96	0.98	~60	0.09

Particle identification

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

Binary classification (separate muons from other particles)

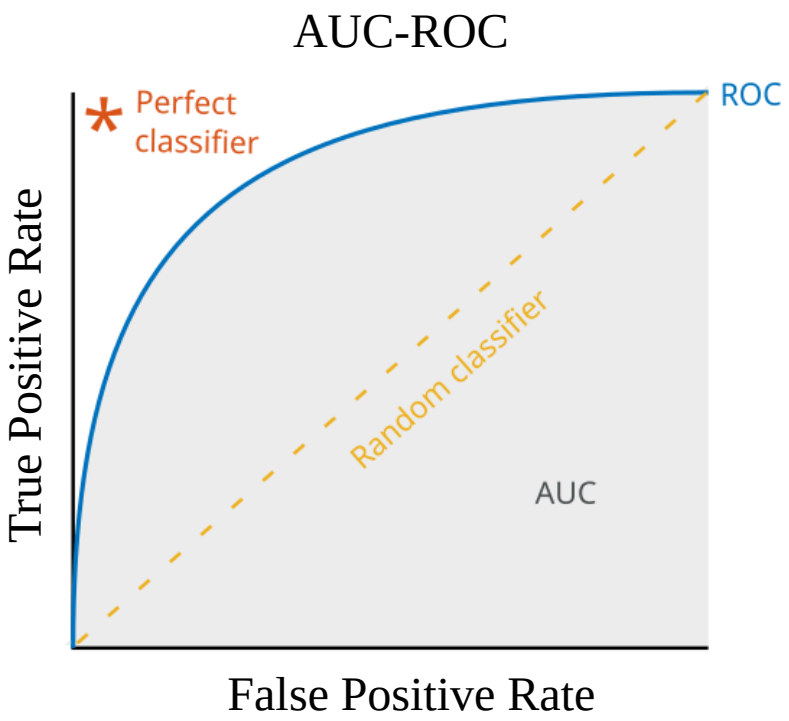
Performance metrics:

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

F1-measure: harmonic mean between the precision and recall, where:

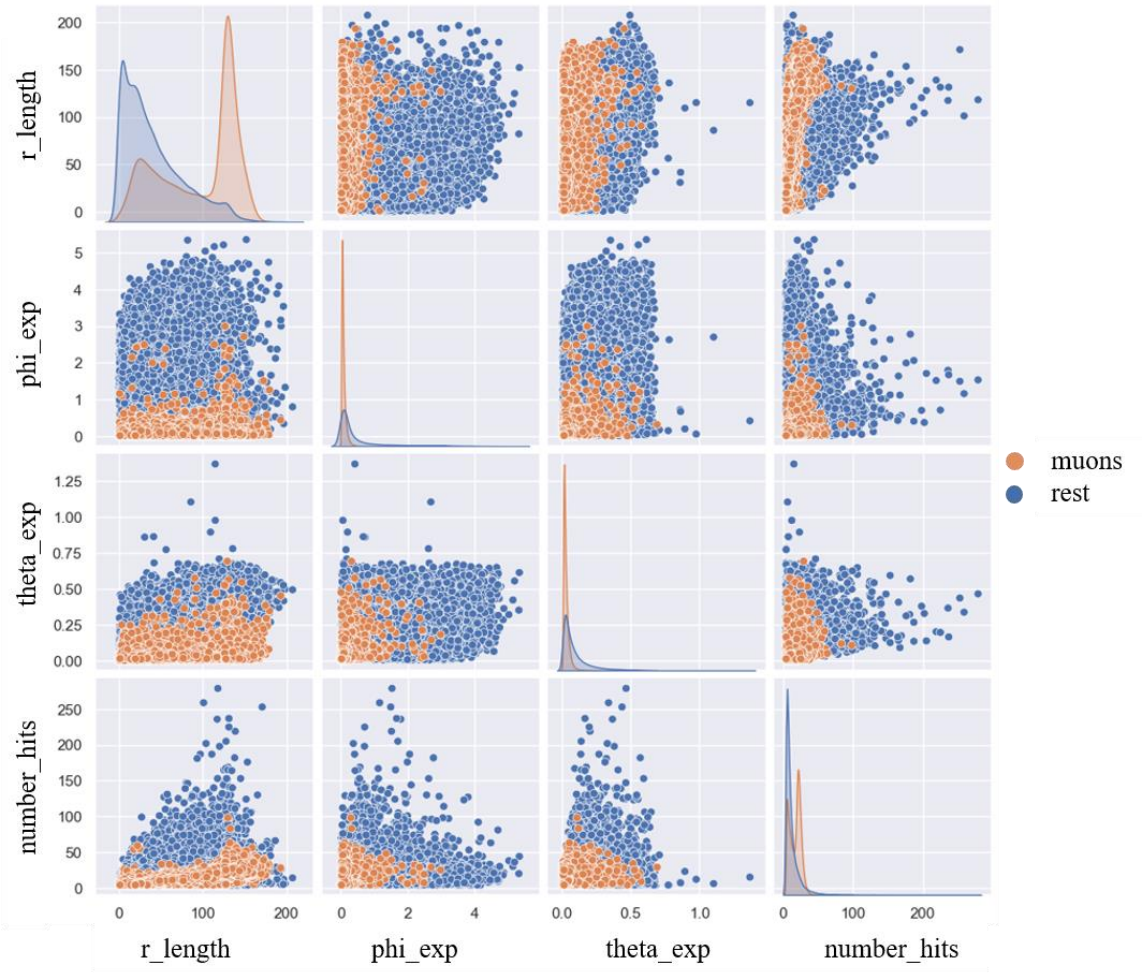
- precision = $\frac{TP}{TP+FP}$
- recall = $\frac{TP}{TP+FN}$

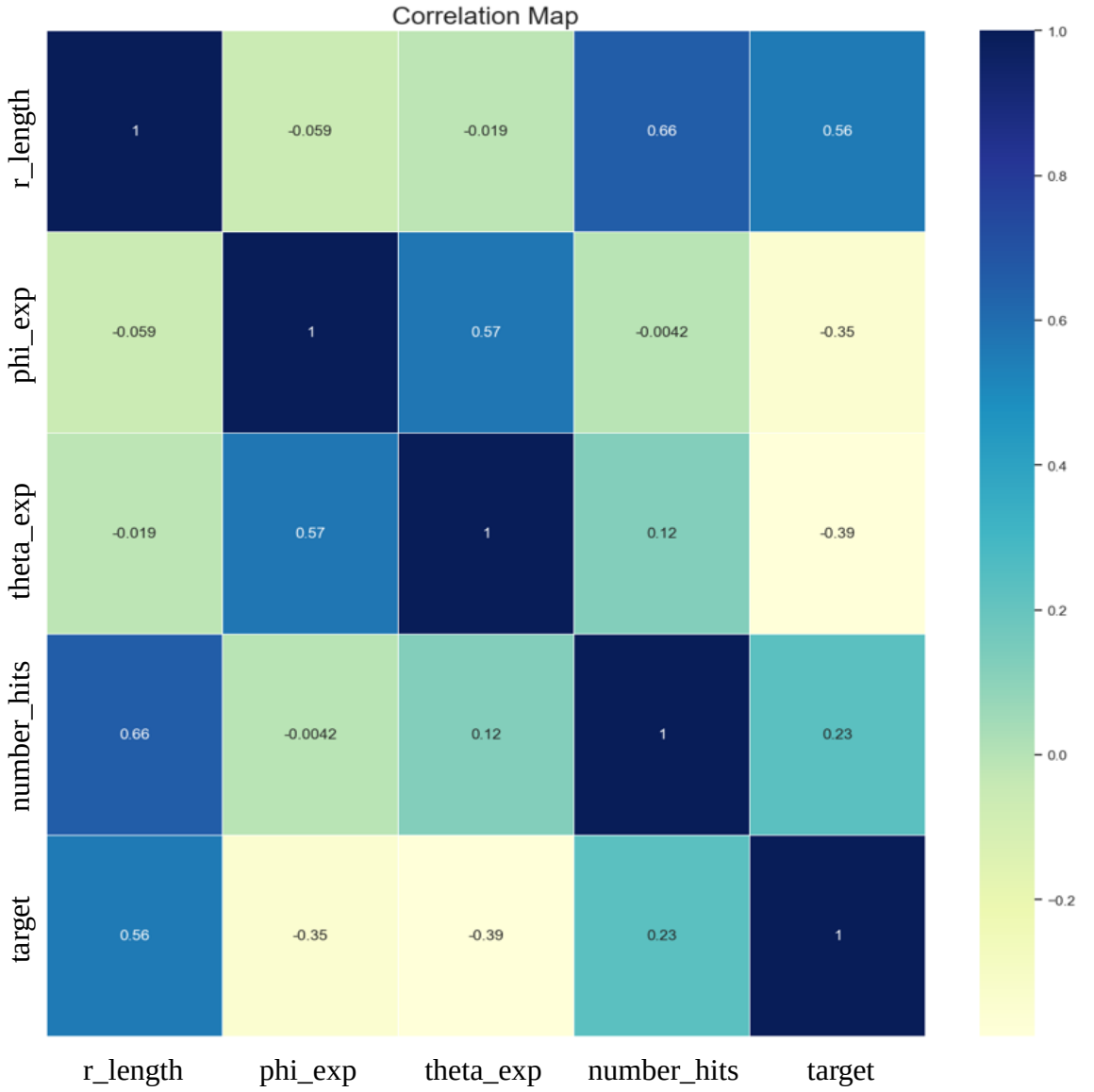
$$\text{f1} = \frac{2*precision*recall}{precision+recall}$$



Features

Feature is an individual measurable property or characteristic of a phenomenon.

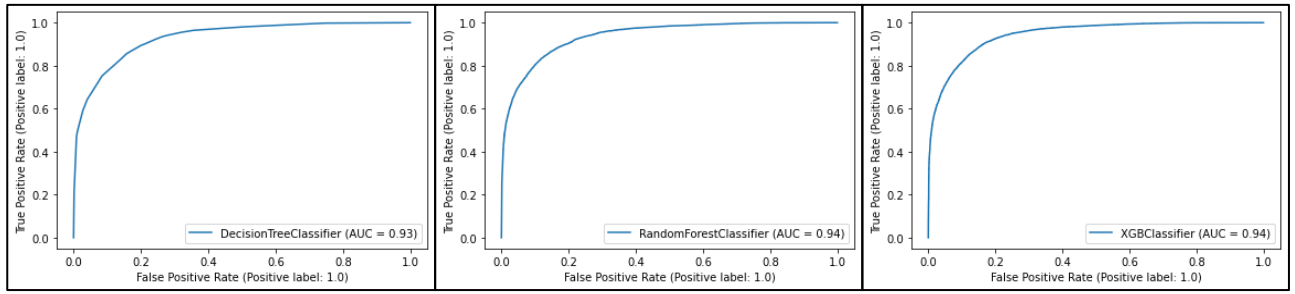




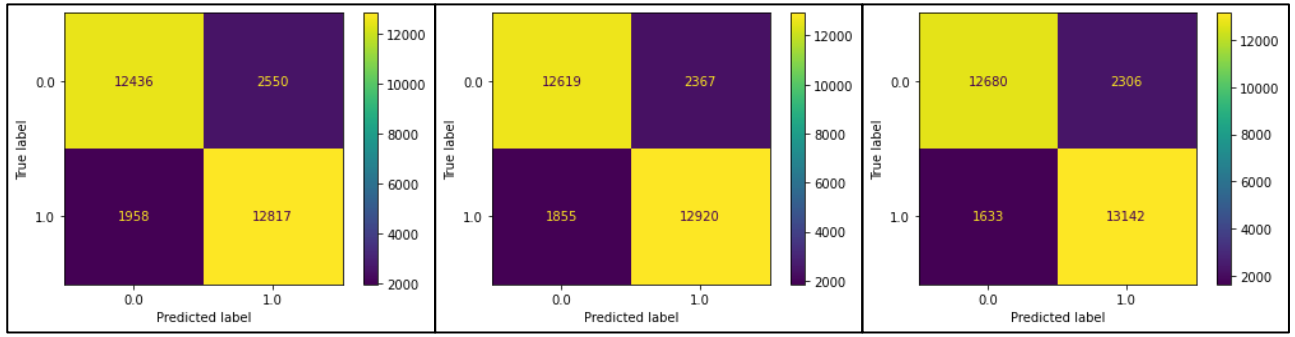
ML algorithms performance

	Decision Tree	Random Forest	XGBoost
Accuracy	0,85	0,86	0,87
F1-score	0,85	0,86	0,87
AUC-ROC	0,93	0,94	0,94

ROC-curve

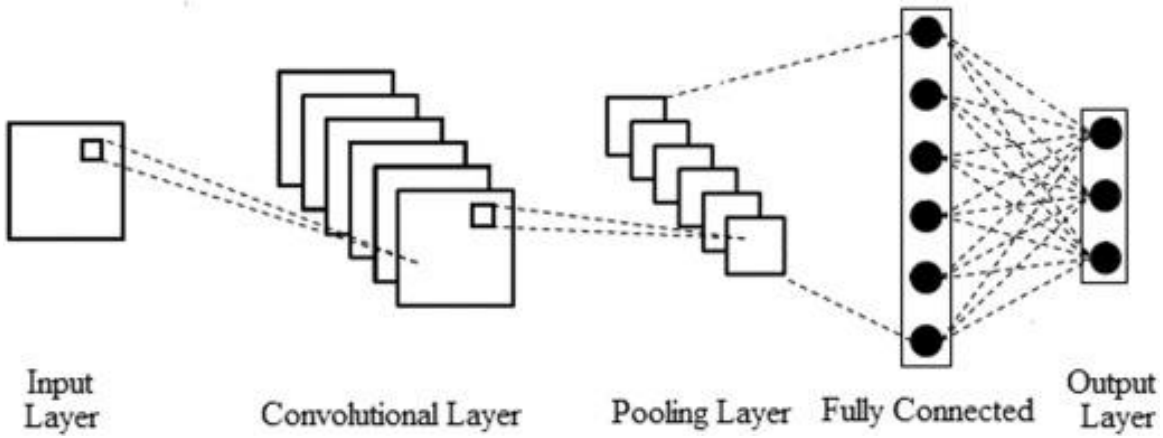


Confusion matrix



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.

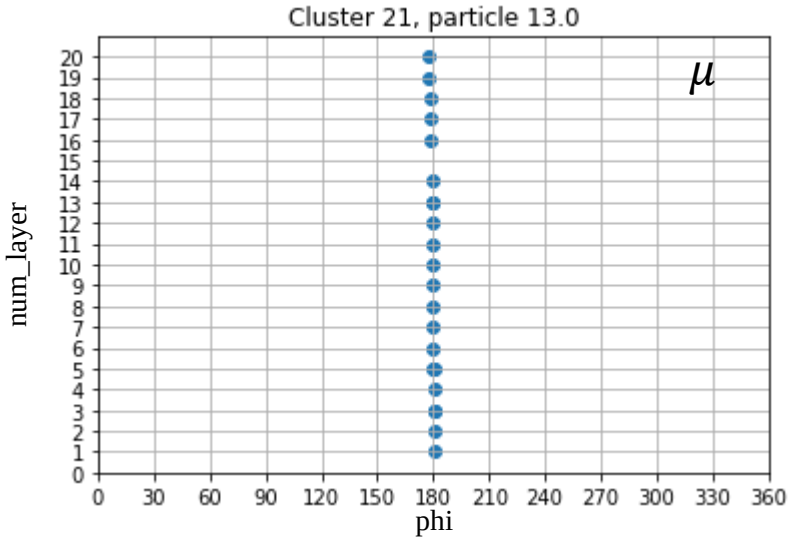
Convolutional layers apply a set of learnable filters to the input image to extract features such as edges, corners, and textures.

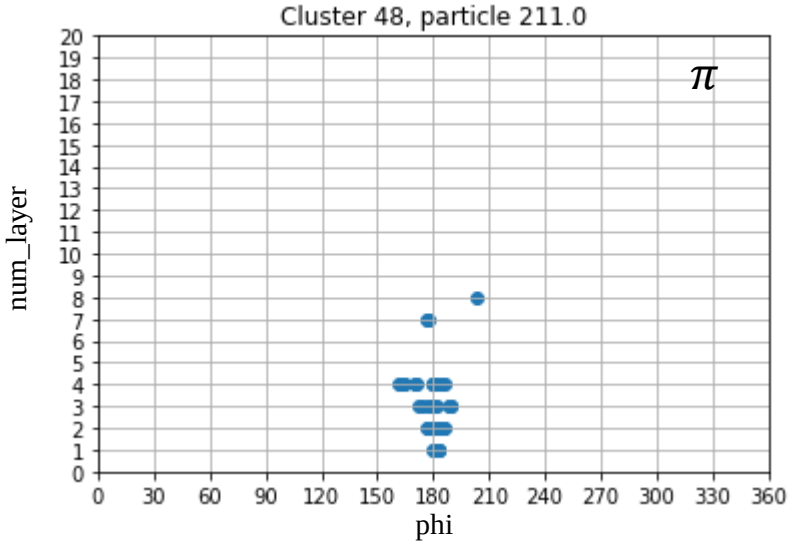
Pooling layers reduce the spatial size of the output of convolutional layers by aggregating nearby values, which helps to reduce the computational cost and overfitting.

Fully connected layers perform the classification or regression task by learning the relationship between the extracted features and the target output.

Convolutional Neural Network

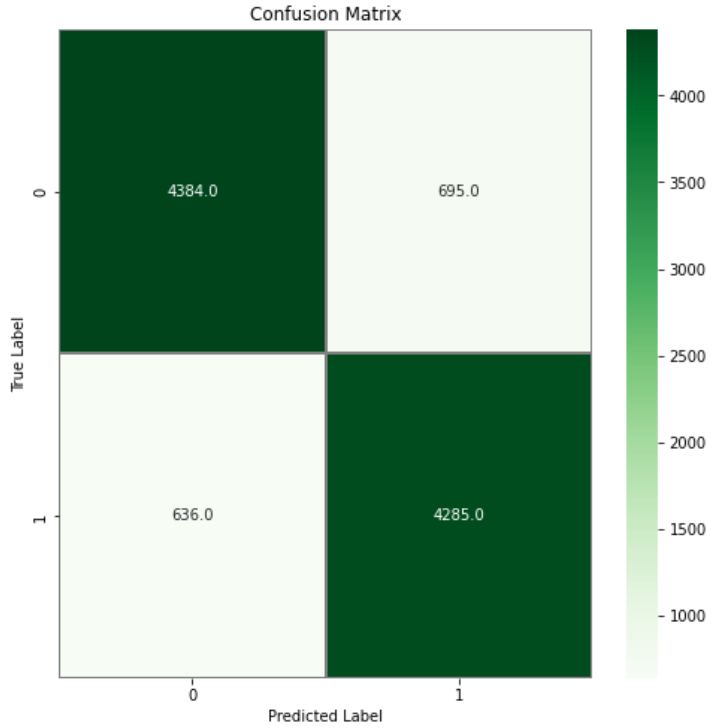
Example of 2d image (20x360) input





Performance

Accuracy	0.87
F1-score	0.86
AUC-ROC	0.94



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 for data filtering we obtained purity of 0.92 and v-measure of 0.94.
3. Decision tree, random forest, XGBoost and convolution neural network were tested as classifiers. In general, all algorithms have shown the similar results (accuracy ~ 0.85-0.87, AUC-ROC ~ 0.93-0.94). However, there is a potential of improving the quality of classification using XGBoost and CNN methods.

Future work:

1. Further improvement of classification ML-methods: using other features, models and their mutual combinations.
2. The use of different configurations of CNNs and analysis of their performance.
3. Multilabel classification: separations of hadrons.
4. Test models on experimental data from the prototype.
5. Compilation of the final pipeline.

Thank you for your attention!