$e^+ - \pi^+$ Electromagnetic Calorimeter shower classification

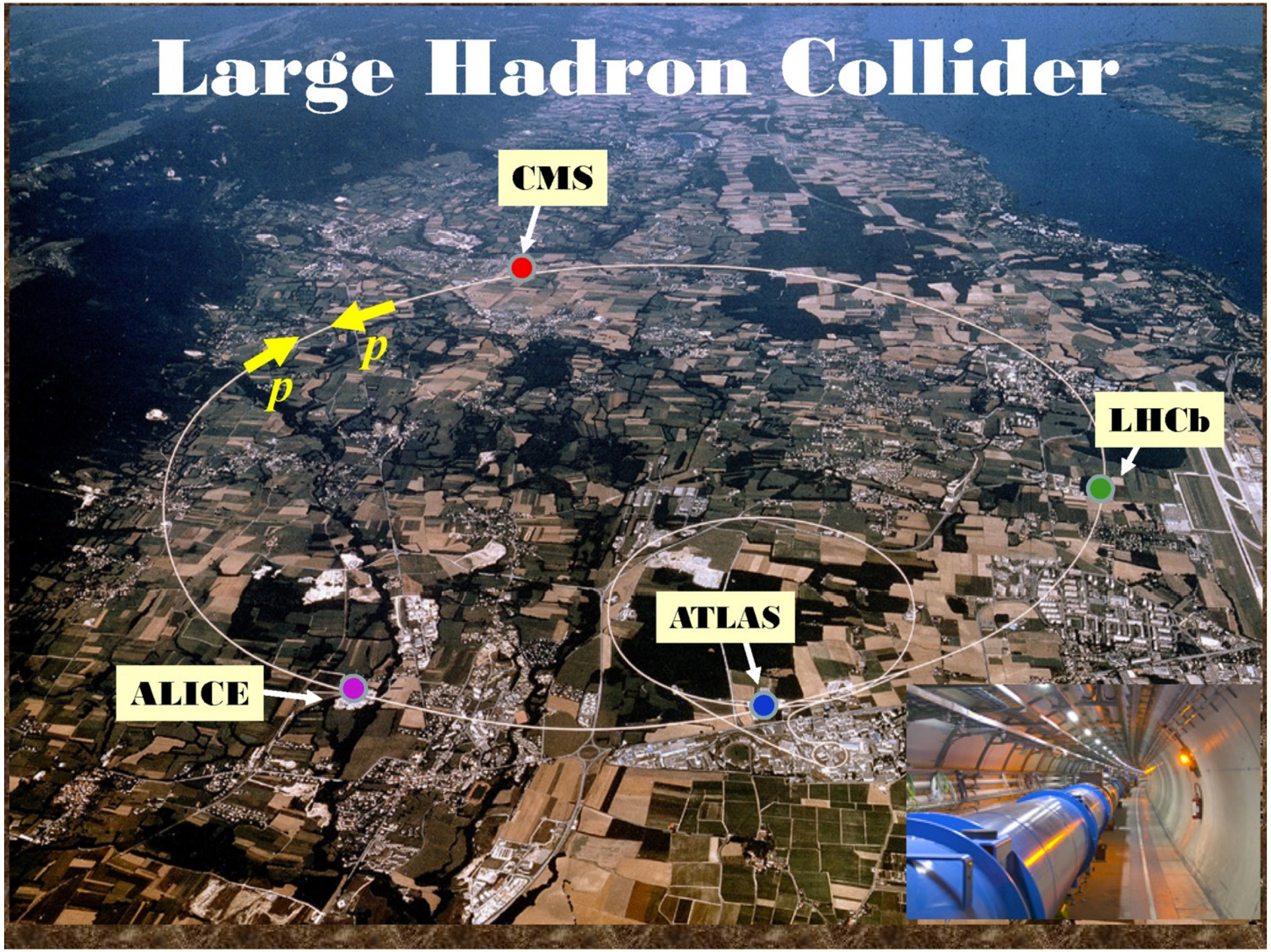
Validation of the machine learning approach on the LHCb use case

Tigran Ramazyan
Eugeny Gurov
Daniil Sobolev

Skoltech

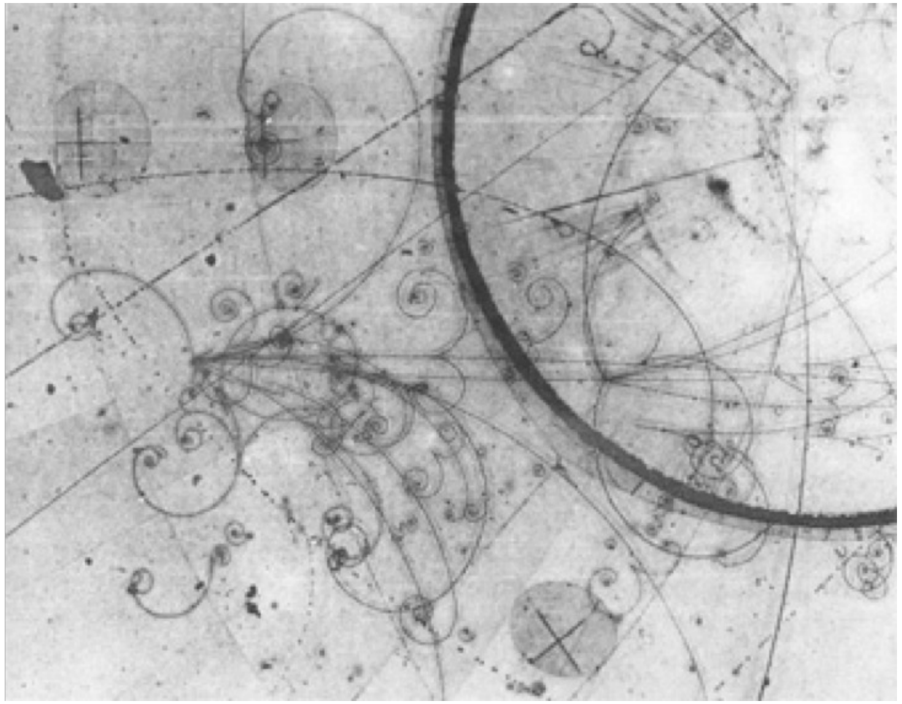
Fall 2022

The LHC

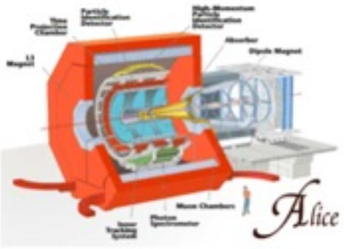


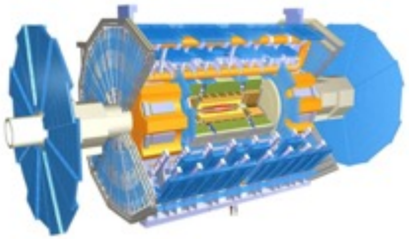
Everything was simpler before…

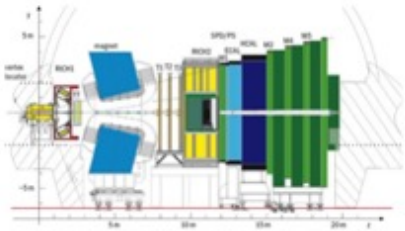
A camera was triggered manually and events were analysed manually.

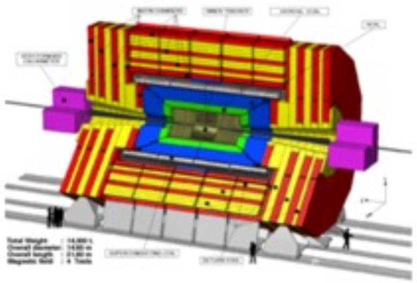


HEP Detectors





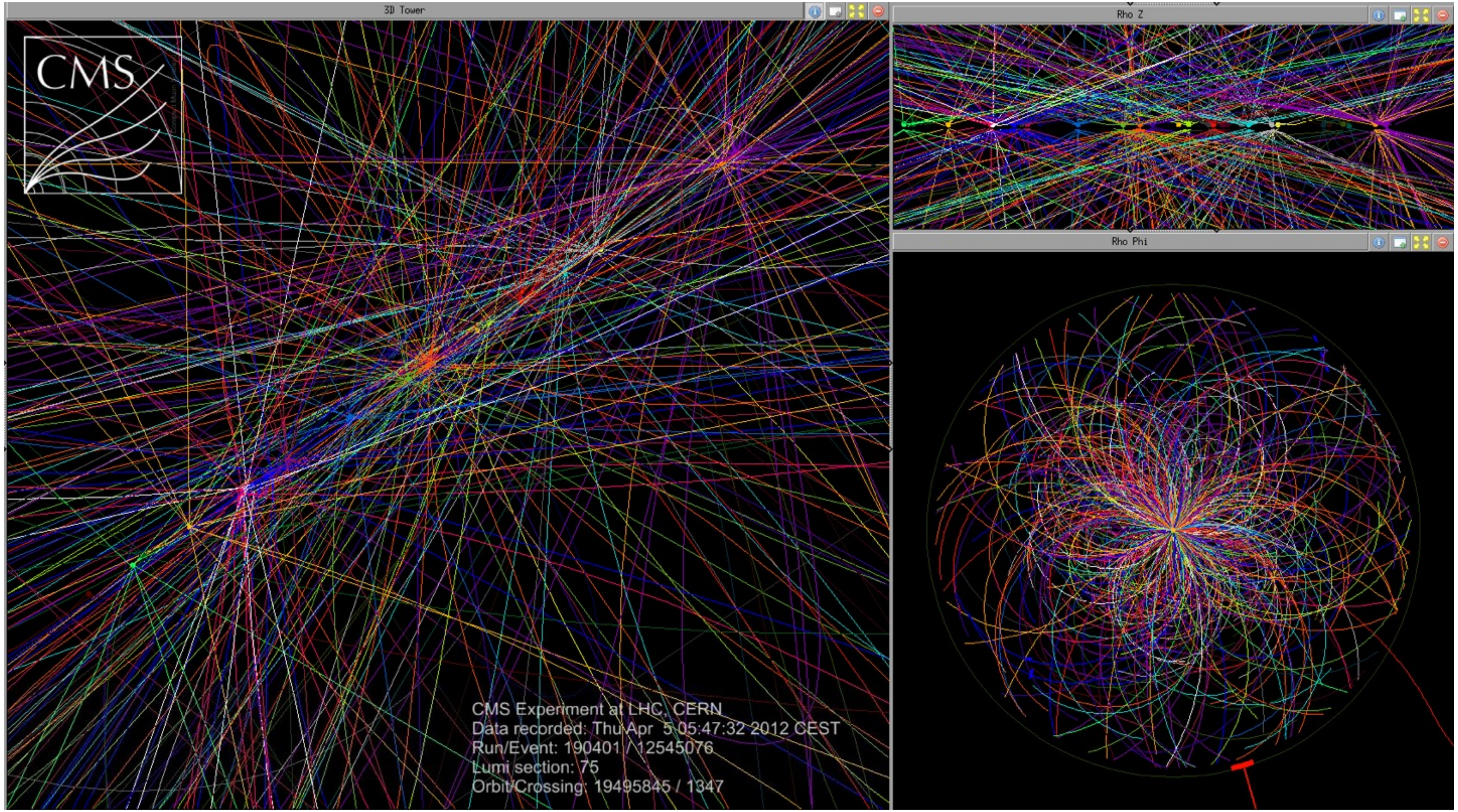




- many layers of sensors: ~200 sq.m. matrices
- resolution: ~100M pixels
- photo speed: 40 000 000 photos per second
- record: 200-1000 photos per second
- work for many years

LHC events today

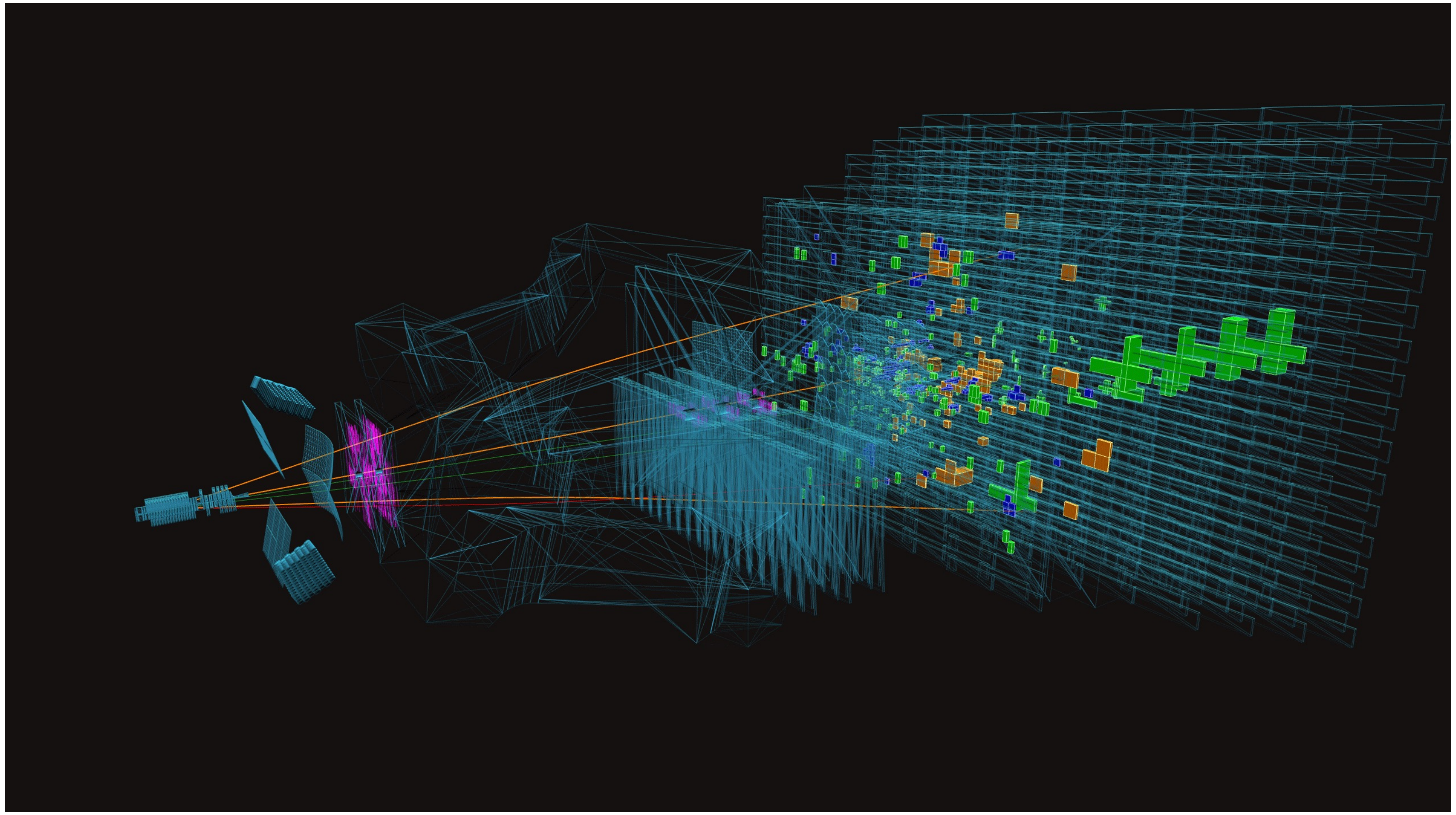
We evidently cannot do without machine analysis, e.g., CV or ML, anymore

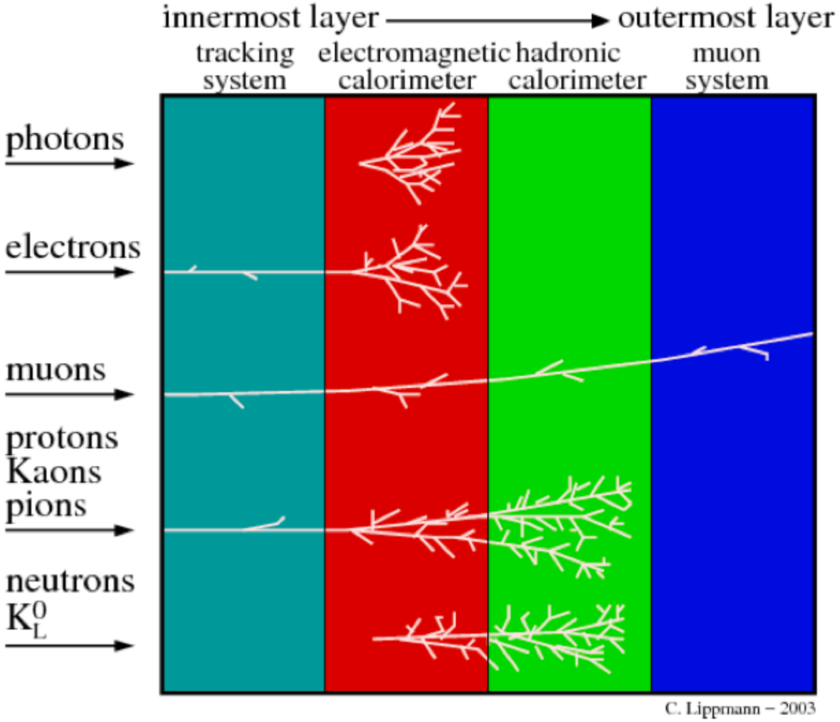


CMS Experiment at LHC, CERN
Data recorded: Thu Apr 5 05:47:32 2012 CEST
Run/Event: 190401 / 12545076
Lumi section: 75
Orbit/Crossing: 19495845 / 1347

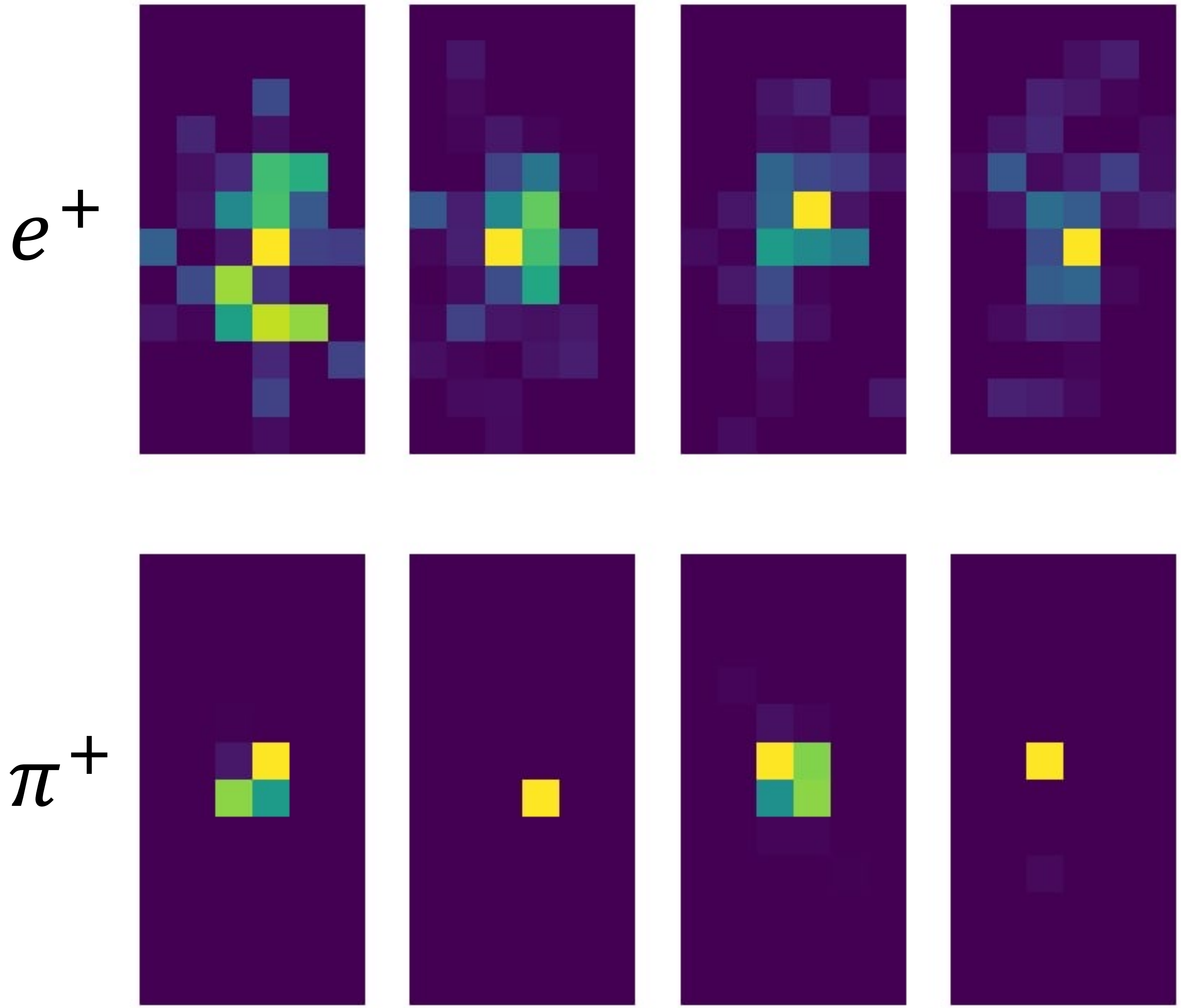
How an event looks like

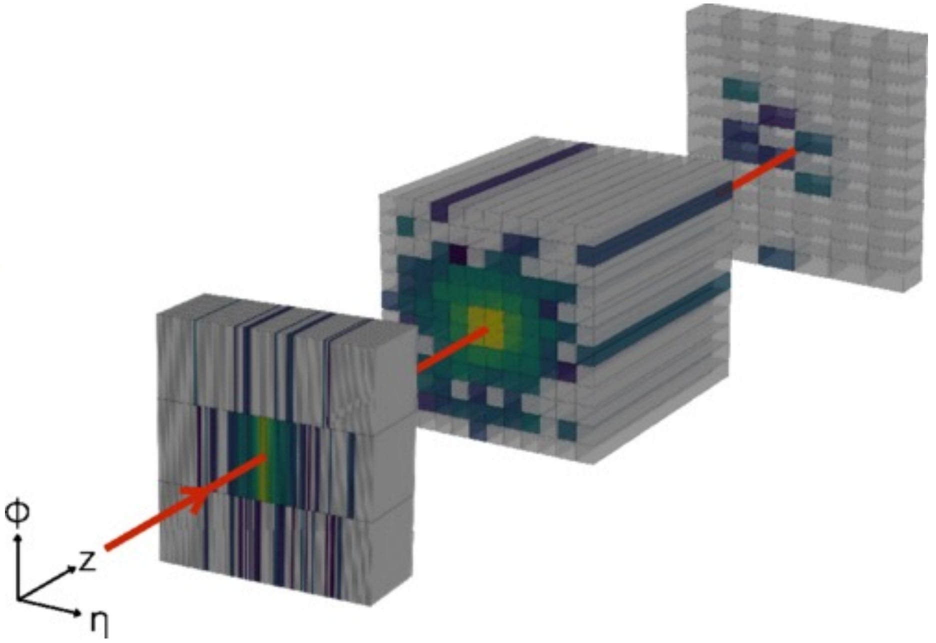
Detectors normally include several subdetectors that react differently on the incident particle.





Examples of data: single calorimeter layer





Expected tasks and results

- Implement CV methods for $e^+ - \pi^+$ classification on calorimeter images
- Compare CV solutions with ML solutions

- Although a CV approach may yield a decent result, it still would be significantly worse than a ML model

Template matching

- Particle showers have unique shapes;
- However shower shapes encapsulate physical characteristics of the particle, e.g., interaction with the electromagnetic field of the detector;
- Averaging over a number (N) of images may give us a template for the corresponding to a particle shower shape;
- N = 100 has shown to yield the best result with **accuracy 0.532**.

N	10	10^2	10^3
Accuracy	0.449	0.532	0.416

	PP	**NP**
P	0.204	0.296
N	0.172	0.328

Confusion matrix for template matching solution.

Thresholding

- Both pions and positrons have highly luminated central pixels;
- Positrons have a significantly large number of luminated pixels outside the central 2x2 pixels;
- We may classify the images based on the number (N) of luminated pixels outside the central 2x2 pixels;
- N = 1 has shown the best result with **accuracy 0.617.**

	PP	**NP**
P	0.467	0.033
N	0.350	0.150

Confusion matrix for thresholding solution.

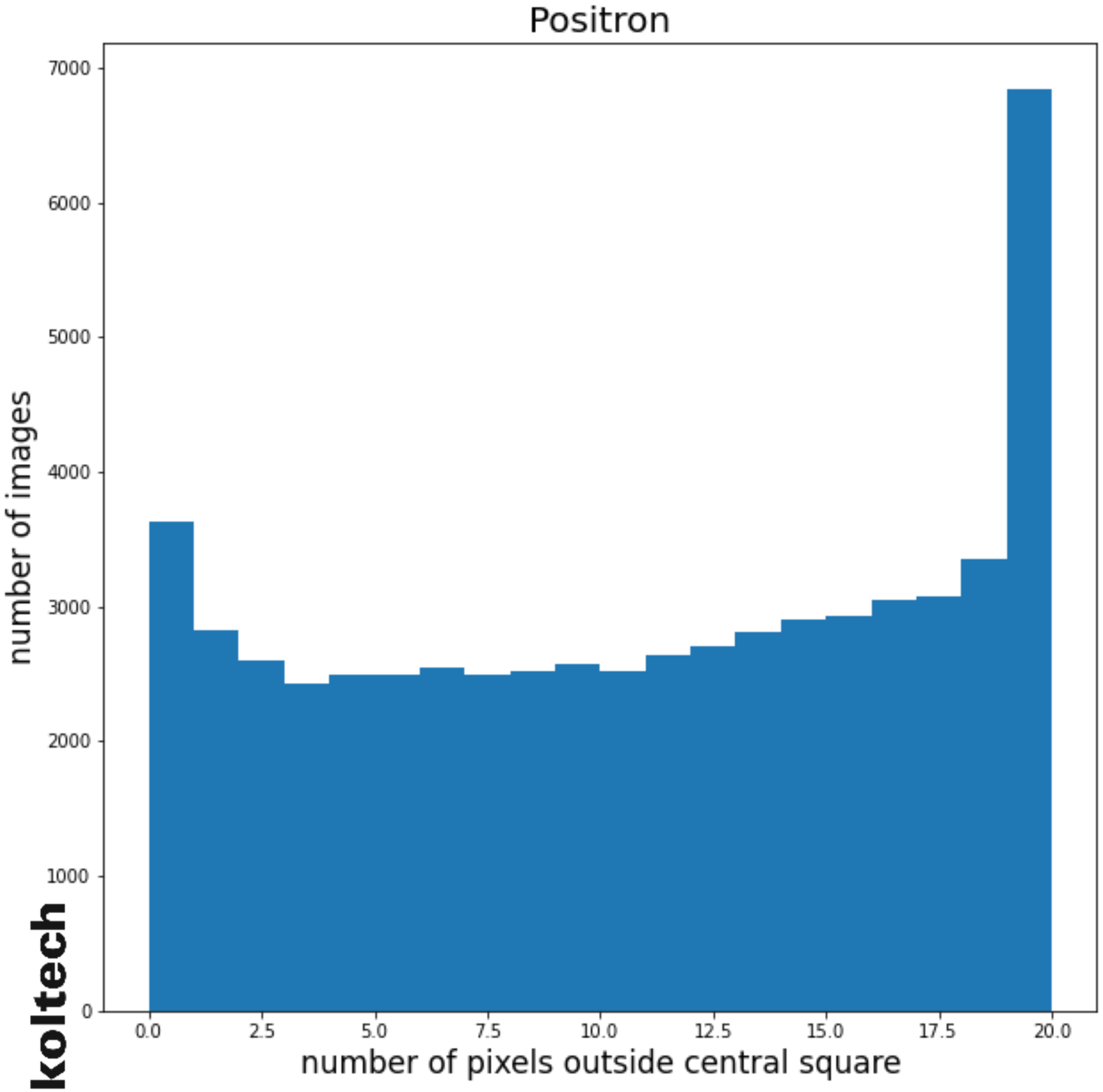
Largest Contour Area Classification

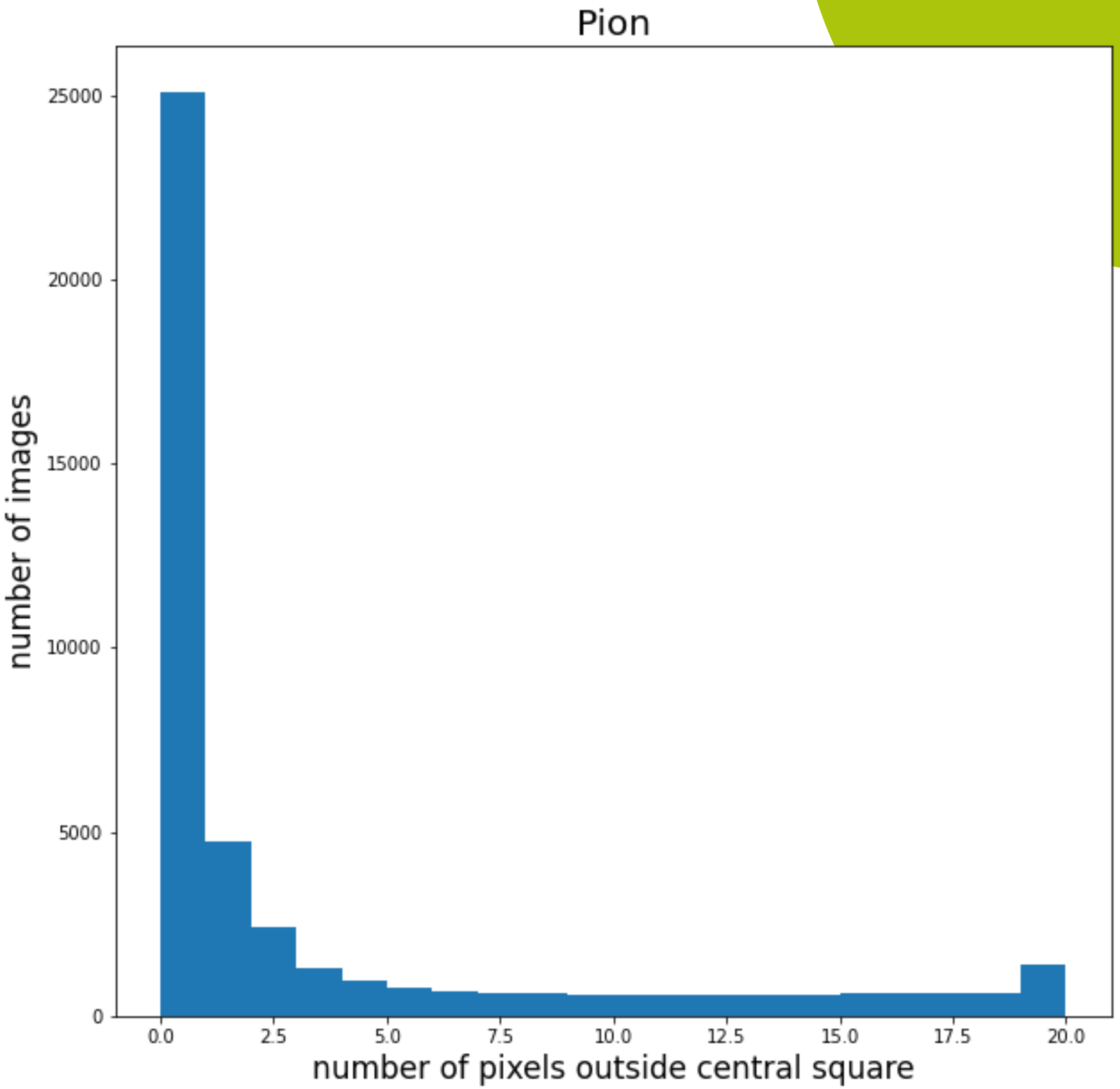
- Continuing the idea from the previous slide, we could classify showers by the area of the largest contour
  - For pions these areas should be at most 4
  - For positrons most of the events should have the largest contour area much greater

Area threshold	3	4	5	6	7	8
Accuracy	0.575	0.562	0.549	0.536	0.522	0.508

	PP	**NP**
P	0.382	0.118
N	0.307	0.193

Confusion matrix for largest contour area classification with area threshold 3.


Positron
number of images
0
1000
2000
3000
4000
5000
6000
7000
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
number of pixels outside central square


Pion
number of images
0
5000
10000
15000
20000
25000
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
number of pixels outside central square

ML Solutions: CNN

- Several CNN architectures have been examined for this task
  - Several (3-4) Conv2d layers followed by (1-2) fully connected layers

- All of them have failed training due to a large amount of non-significant values (zero pixels) and have yield ‘random classifier’ results, i.e., accuracy of ~0.5
  - Transformations such as CenterCrop have not make any significant improvement

- However, the method currently in production is a DenseNet-based NN with 99.99% accuracy.

ML Solutions: CatBoost

- “Gradient boosting is also utilized in High Energy Physics in data analysis. At the Large Hadron Collider (LHC), variants of gradient boosting Deep Neural Networks (DNN) were successful in reproducing the results of non-machine learning methods of analysis on datasets used to discover the Higgs boson.” [1]

- CatBoost handles well low dimensional data (images are 12x6, i.e., 72x1 vectors)

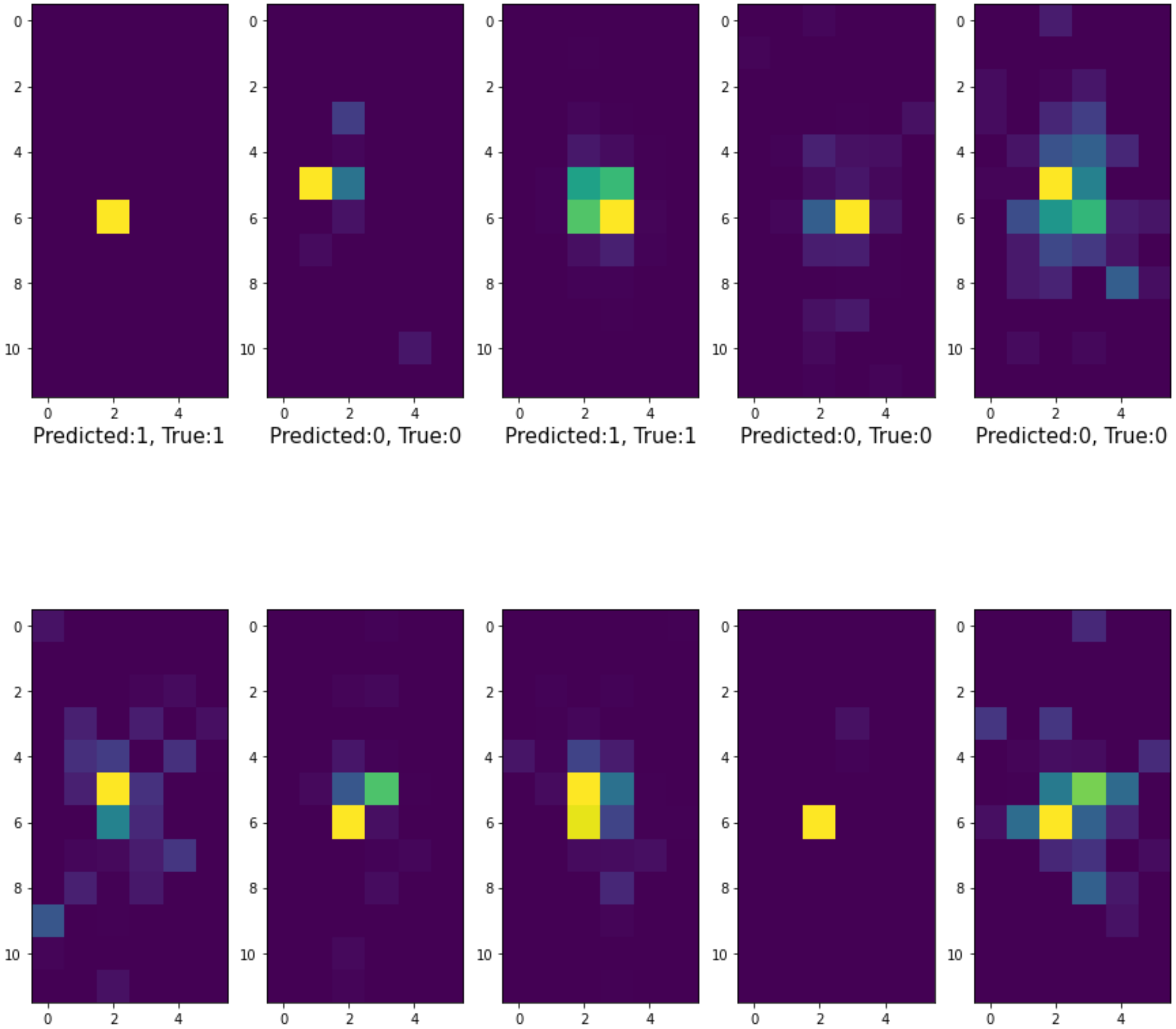
- The model has shown the best result with **accuracy** of **0.985**

	PP	**NP**
P	0.496	0.004
N	0.011	0.489

Confusion matrix for CatBoost.

[1] https://en.wikipedia.org/wiki/Gradient_boosting

CatBoost well predicted example



Predicted:1, True:1

Predicted:0, True:0

Predicted:1, True:1

Predicted:0, True:0

Predicted:0, True:0

Predicted:0, True:0

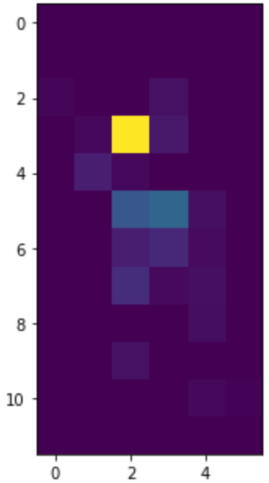
Predicted:1, True:1

Predicted:1, True:1

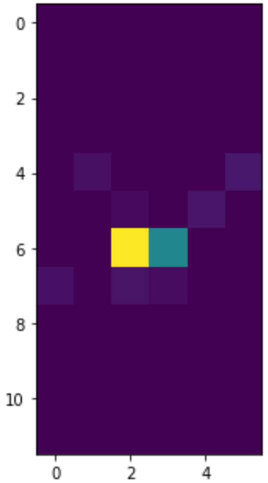
Predicted:1, True:1

Predicted:0, True:0

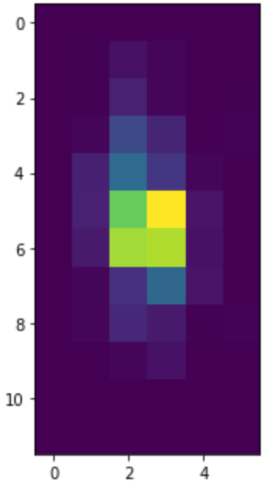
CatBoost misprediction examples



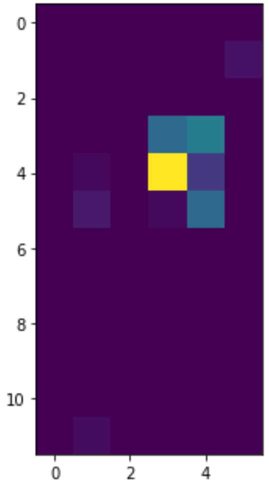
Predicted:0, True:1



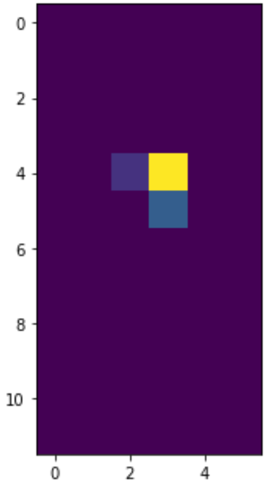
Predicted:1, True:0



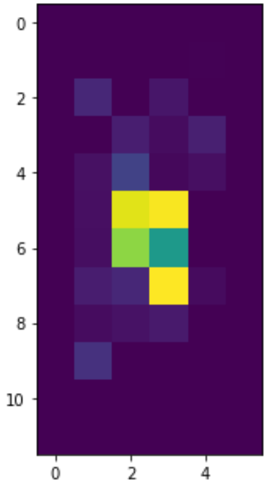
Predicted:0, True:1



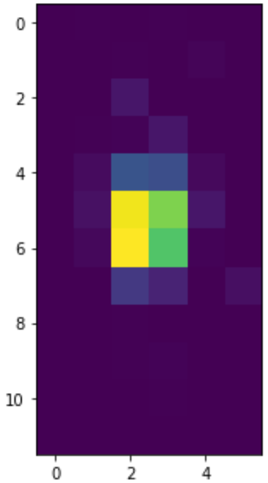
Predicted:0, True:1



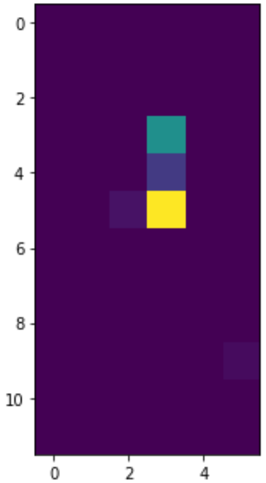
Predicted:0, True:1



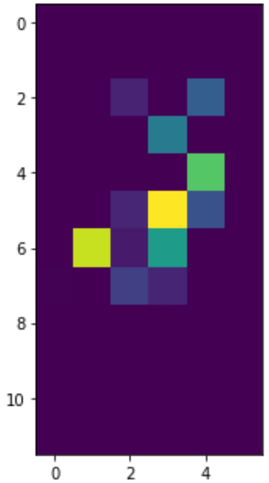
Predicted:0, True:1



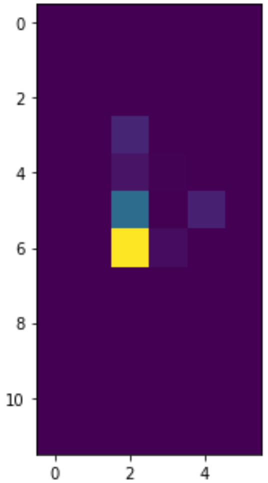
Predicted:0, True:1



Predicted:1, True:0



Predicted:0, True:1



Predicted:0, True:1

Conclusion

- Several CV methods have been tested as solution for the $e^+ - \pi^+$ electromagnetic calorimeter shower classification.
- Although they may be extended into more precise solutions, a plug-and-play GB solution yields a result close to the state-of-the-art model.