Применение машинного обучения в задаче идентификации частиц

В. Папоян[1]

Соавторы: А. Айриян[1], А. Апарин[2], О. Григорян[1], А. Мудрох[2], А. Коробицин[2]

[1]МЛИТ ОИЯИ, [2]ЛФВЭ ОИЯИ

17.11.2022

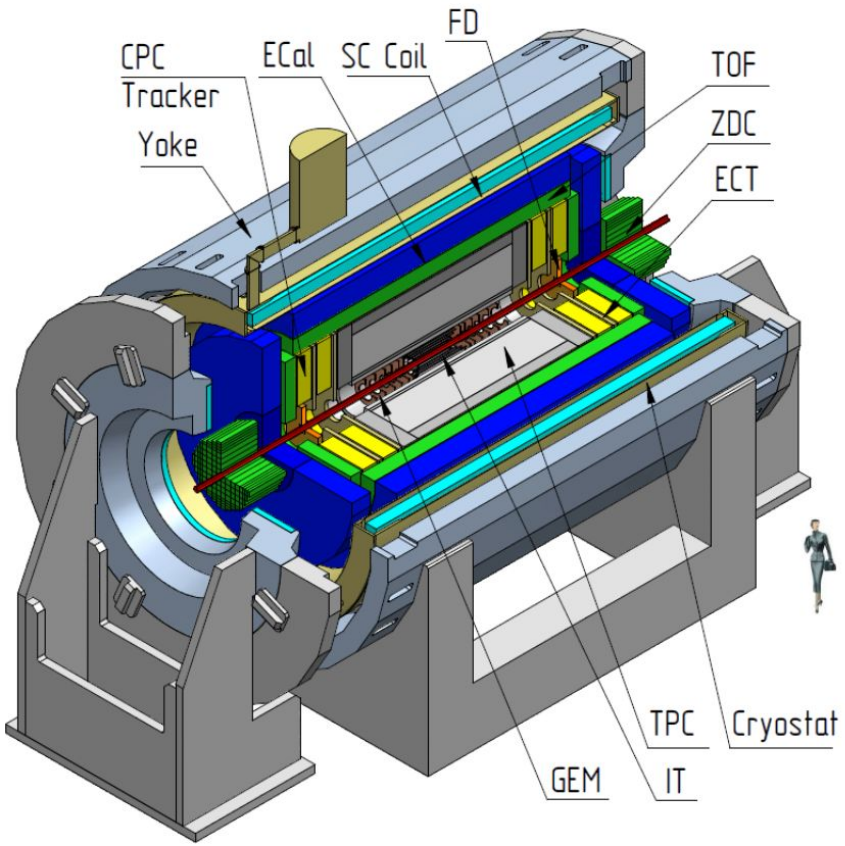
Идентификация частиц в MPD

Идентификация частиц в MPD основана на детекторах **TPC** и **TOF**.

Время Проекционная Камера (TPC) - газовый детектор с считывающими устройствами на торцах. Измеряет трехмерные координаты пролетающих частиц [1].

Времяпролетная система (ToF) определяет скорость полета частицы посредством измерения времени полета частицы от точки взаимодействия до самой ToF.

Детектор MPD:



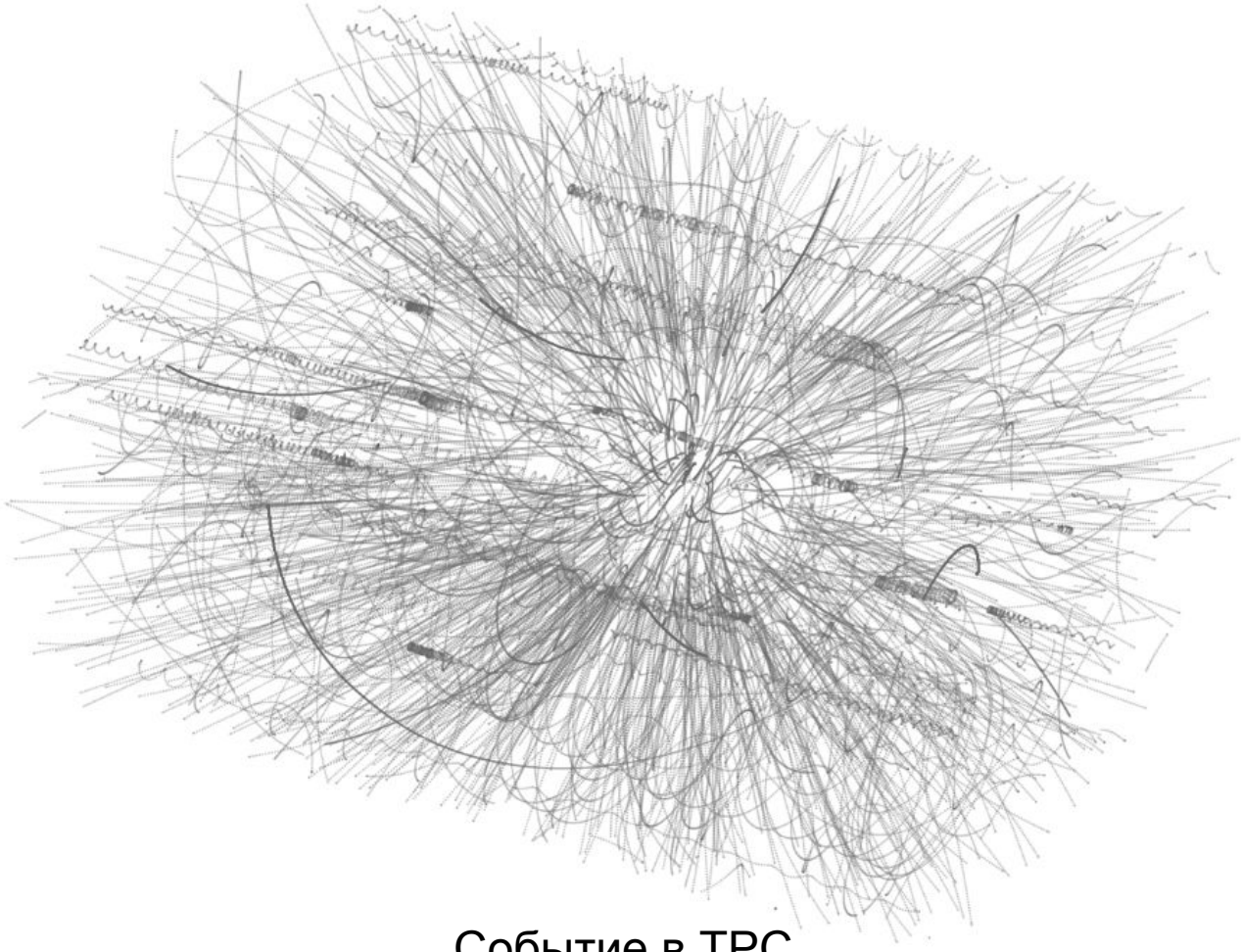
[1] Hilke H. J. Time projection chambers //Reports on Progress in Physics. – 2010. – Т. 73. – №. 11. – С. 116201.

Идентификация частиц в MPD

Идентификация частиц (**PID**) это задача определения типа частицы по заданным характеристикам ее трека.

Для идентификации частицы необходима информация о ее:

- **импульсе**
- **заряде**
- **удельной потере энергии**
- **квадрате массы** (TPC + TOF)



Событие в TPC

Классификация частиц

Идентификация частиц является задачей **классификации** в машинном обучении (обучение **с учителем**).

Пусть

X - множество описаний объектов (характеристик частиц)

Y - конечное множество меток классов

Существует **неизвестная** целевая зависимость - отображение

$$y^* : X \to Y,$$

значения которой известны только на объектах конечной обучающей выборки

$$X^m = (x_1, y_1), \ldots, (x_m, y_m),$$

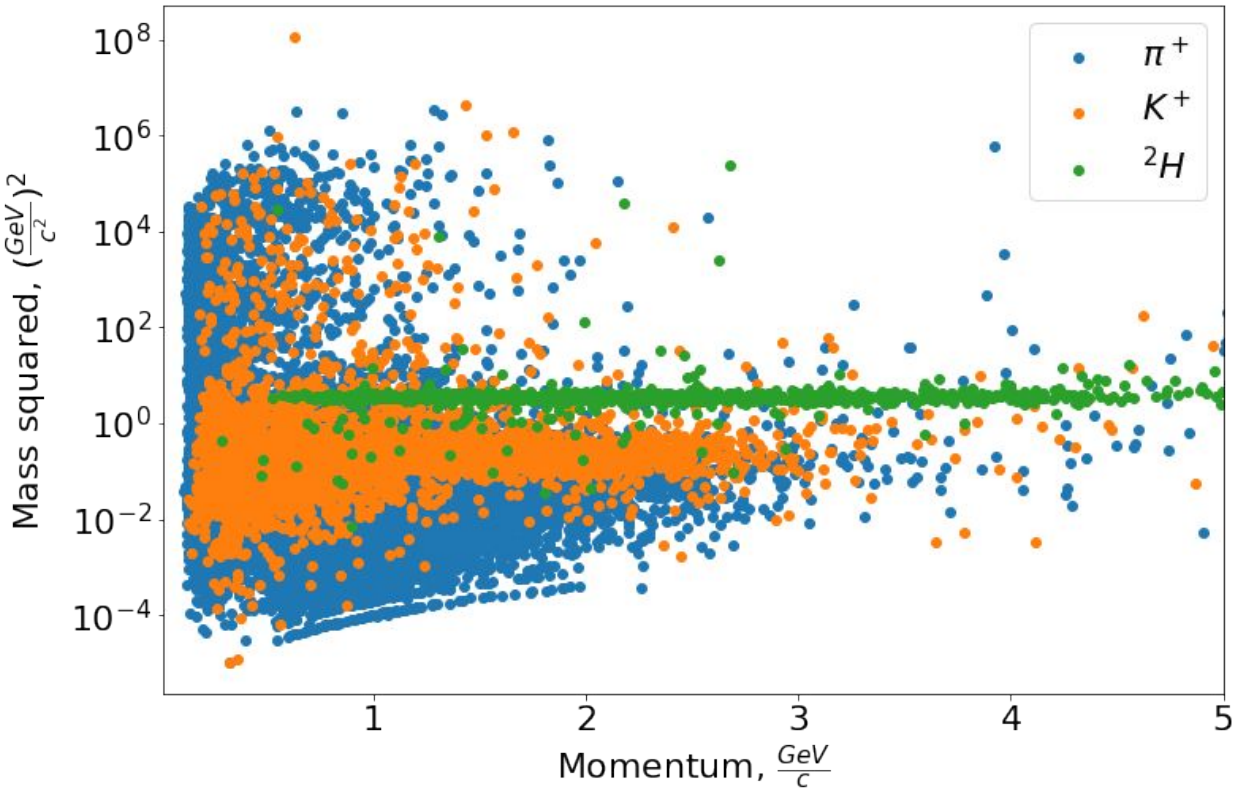
Требуется построить алгоритм **a** способный классифицировать произвольный объект $x \in X$

$$a : X \to Y.$$

Классификация частиц

Задача идентификации частиц может быть рассмотрена как:

1. Мультиклассификация, когда все частицы отделяются друг от друга одновременно;
2. Бинарная классификация:
   a. one-vs-rest;
   b. one-vs-one.



Наборы данных

prod01 (реальное распределение частиц):

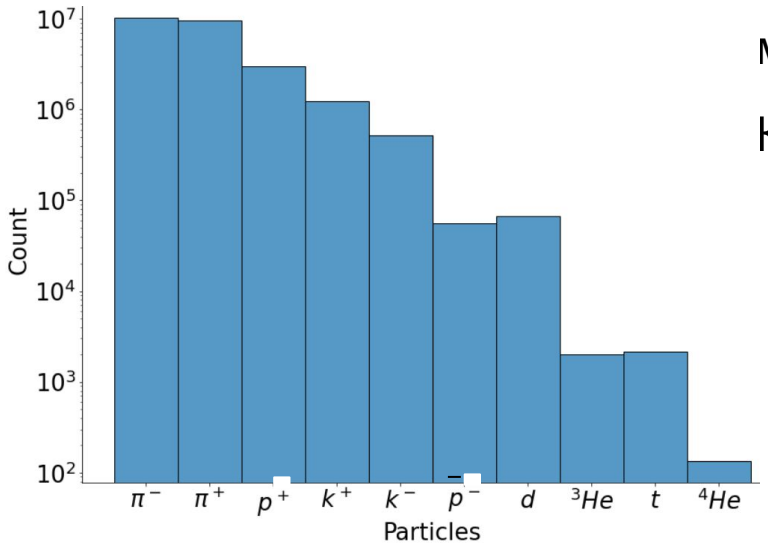
PHQMD

Geant 4 minimum bias

Bi+Bi @ 9.2 GeV

full detector configuration

Количество треков: 24M



prod04 (искусственное добавление частиц):

URQMD + BOX;

Geant 4 minimum bias

Bi+Bi @ 9.2 GeV

full detector configuration + SmearVertexXY (неопределенность места вершины) 1.1 cm

Количество треков: 4.5M

prod05 (реальное распределение частиц):

Request 25 URQMD;

Geant 4 minimum bias

Bi+Bi @ 9.2 GeV

full detector configuration

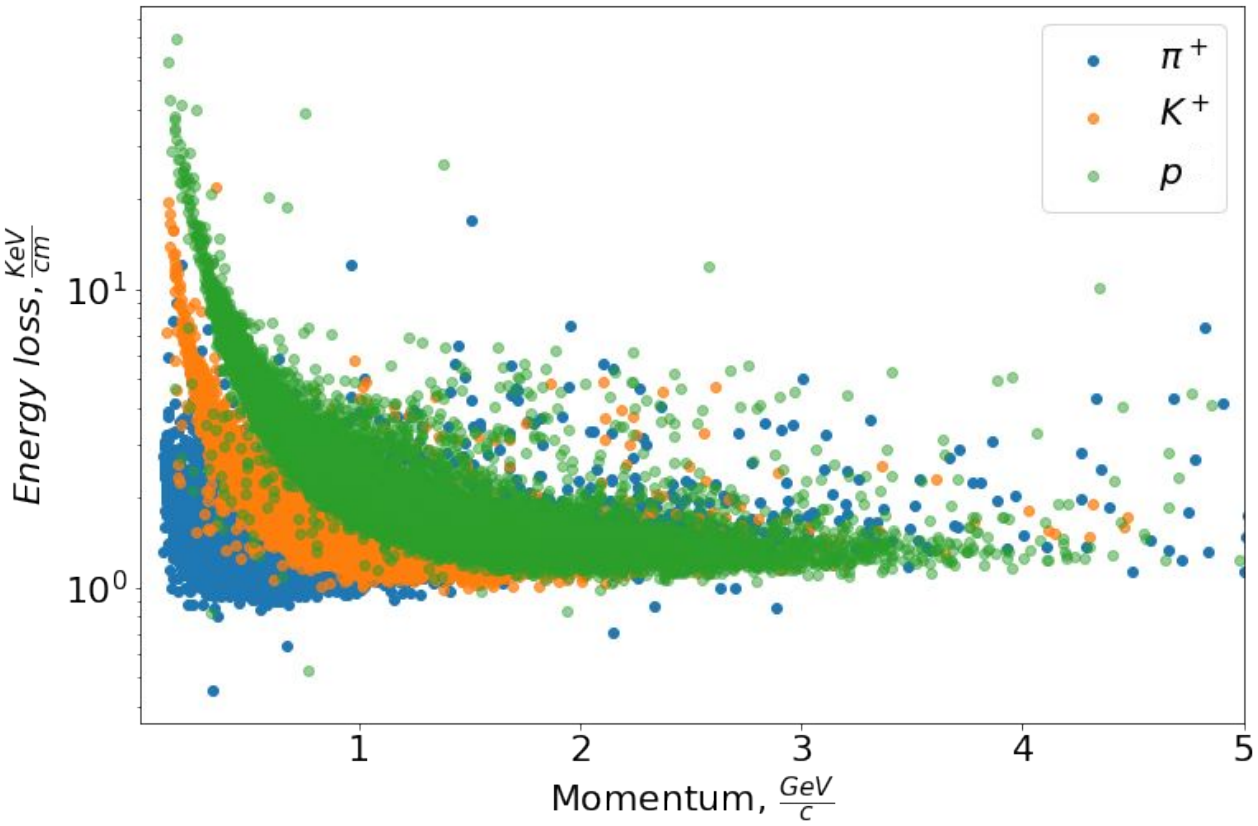
Количество треков: 5.8M

Данные для **тестирования** сгенерированы при тех же условиях, что и prod05 (2.6 M)

Наборы данных

Рассматривались 6 классов частиц:

Протон (p); Каоны (K^+ , K^-); Пионы (π^+ , π^-); Антипротон ($\bar{p}$).

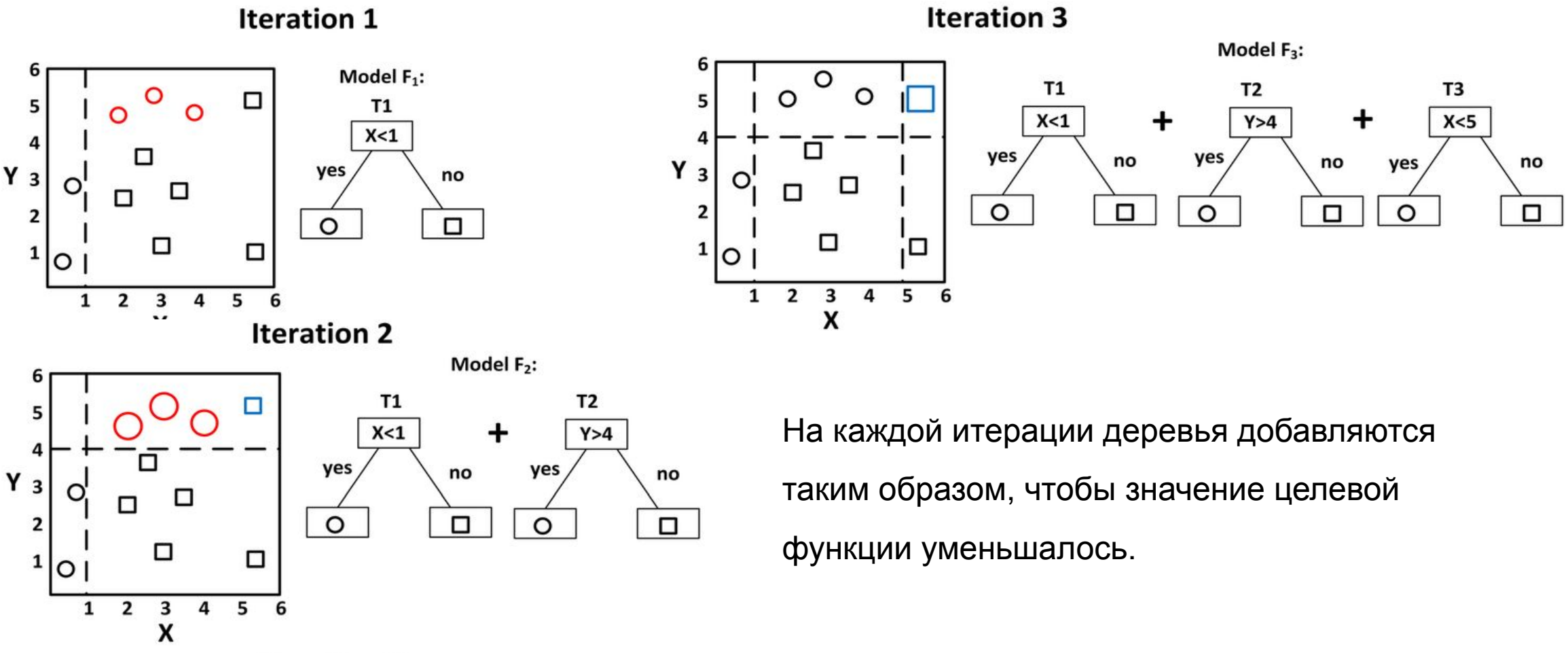


Вектор признаков:

- импульс (p)
- заряд (q)
- потеря энергии (de/dx)
- квадрат массы (m^2)
- количество откликов частицы в TPC (nHits)
- псевдобыстрота (eta)
- вершина взаимодействия (Vx, Vy, Vz)

Бустинг деревьев решений (CatBoost)

Градиентный бустинг - алгоритм, строящий ансамбль решающих деревьев небольшой глубины.



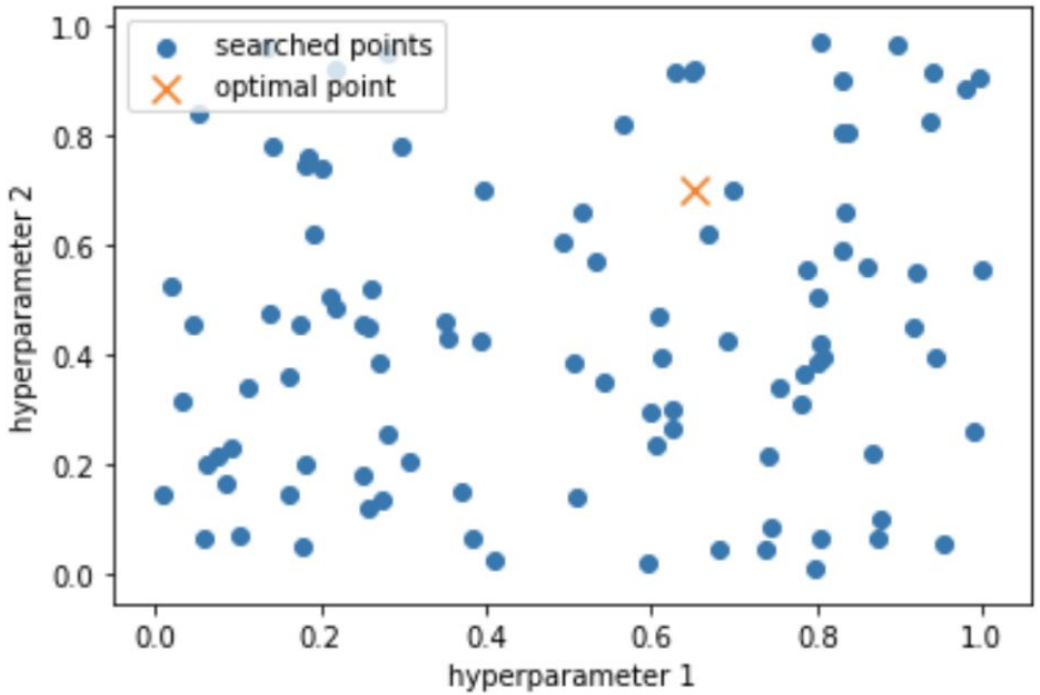
На каждой итерации деревья добавляются таким образом, чтобы значение целевой функции уменьшалось.

https://www.researchgate.net/figure/A-simple-example-of-visualizing-gradient-boosting_fig5_326379229

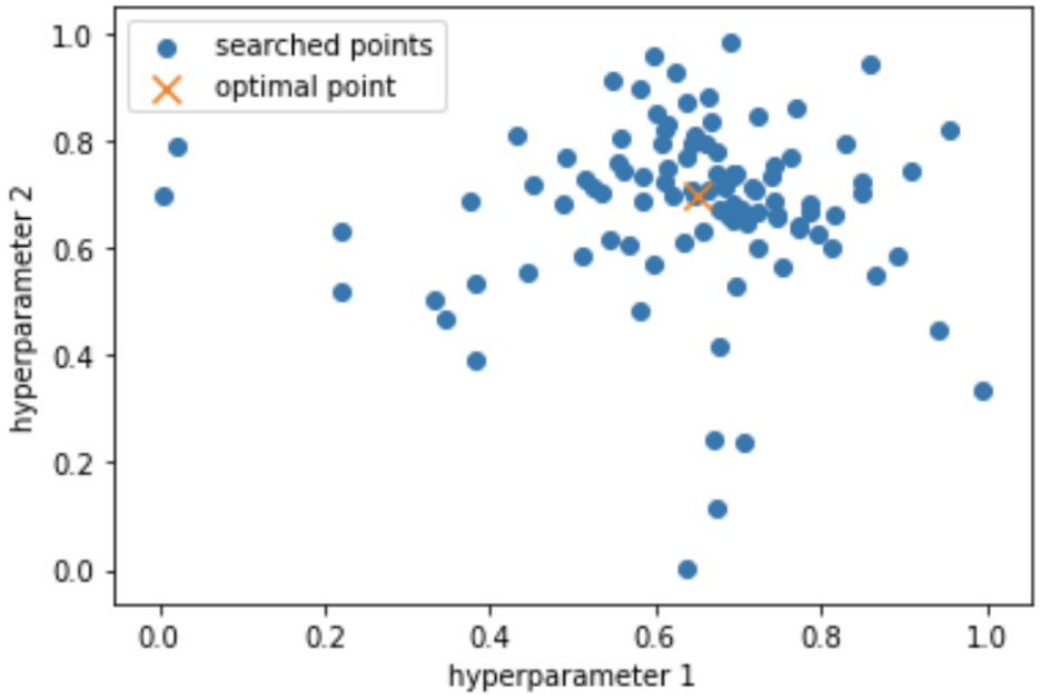
Подбор оптимальных гиперпараметров модели

Поиск оптимальных гиперпараметров был осуществлен с помощью алгоритма Tree-structured Parzen Estimater (TPE), который является разновидностью байесовской оптимизации.

Случайный поиск



Поиск методом TPE

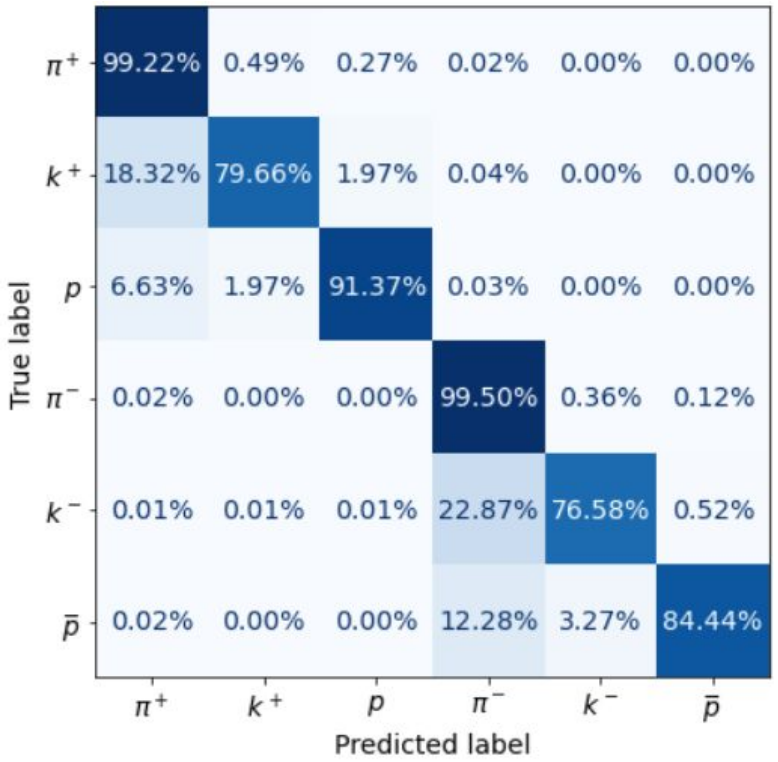


https://miro.medium.com/max/4800/1*tYWqO5BwNDVaM3kP3w1IAg.png

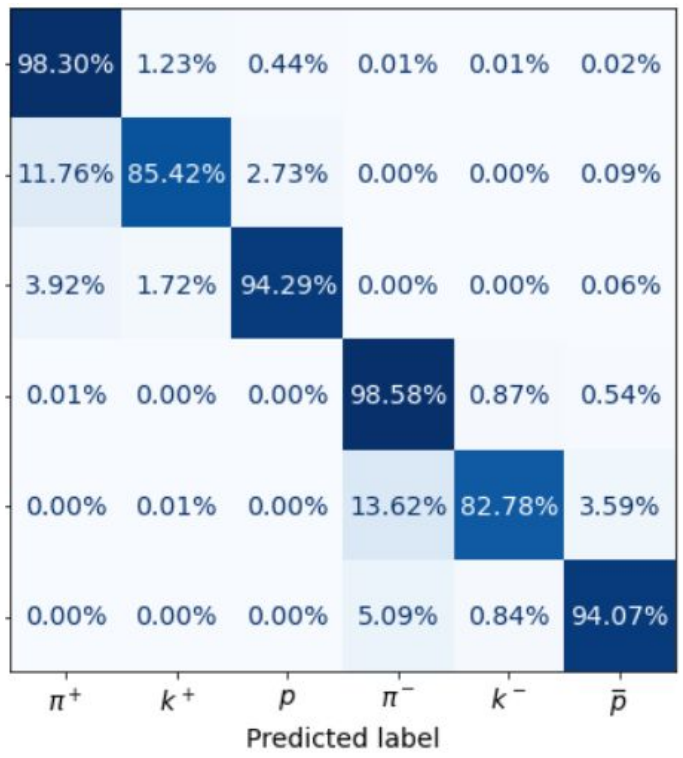
Матрицы ошибок

Каждая столбец матрицы - предсказанное моделью значение класса, а каждая строка - фактическое значение

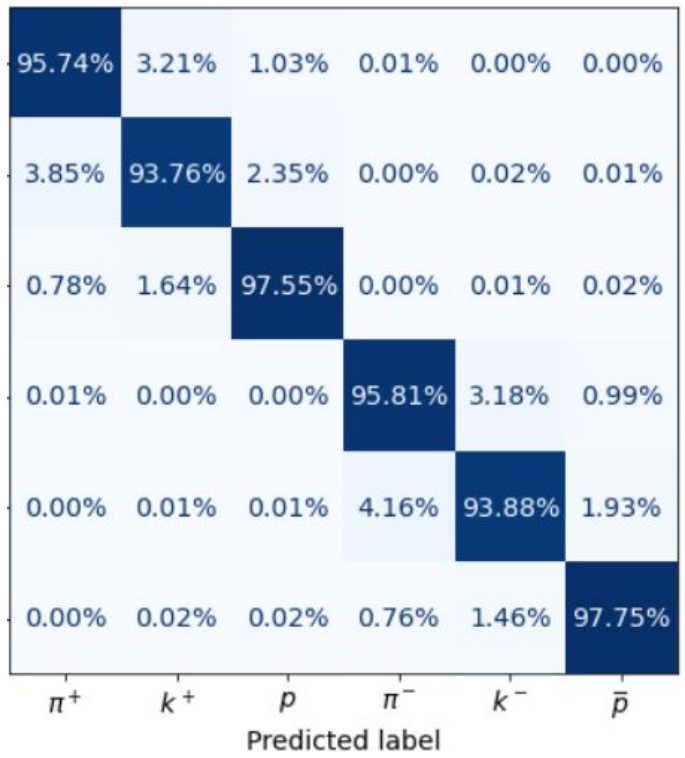
prod01:



prod04:



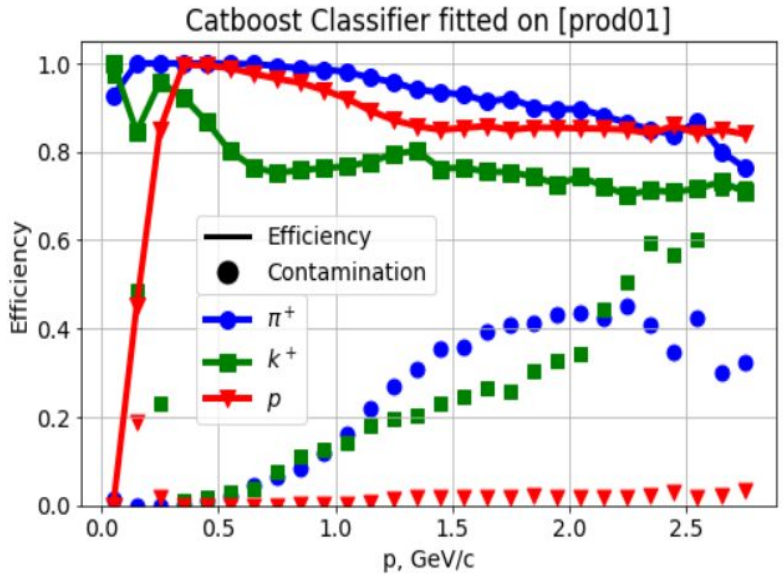
prod05:

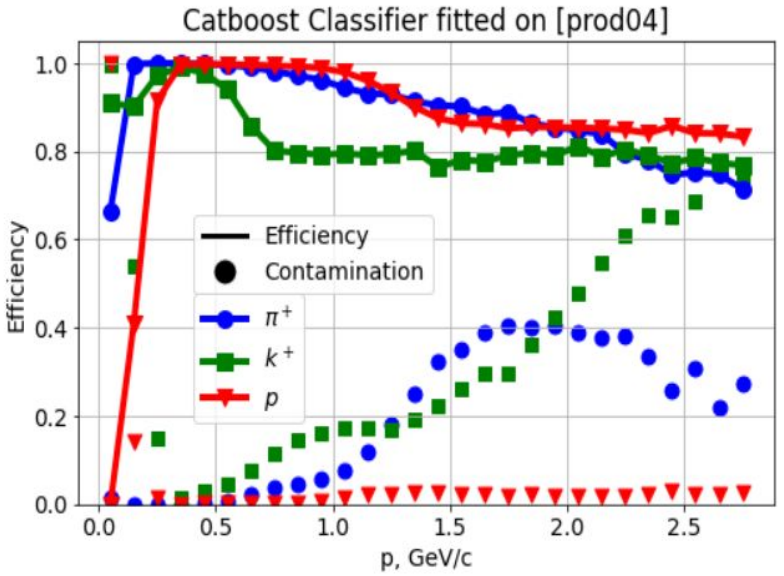


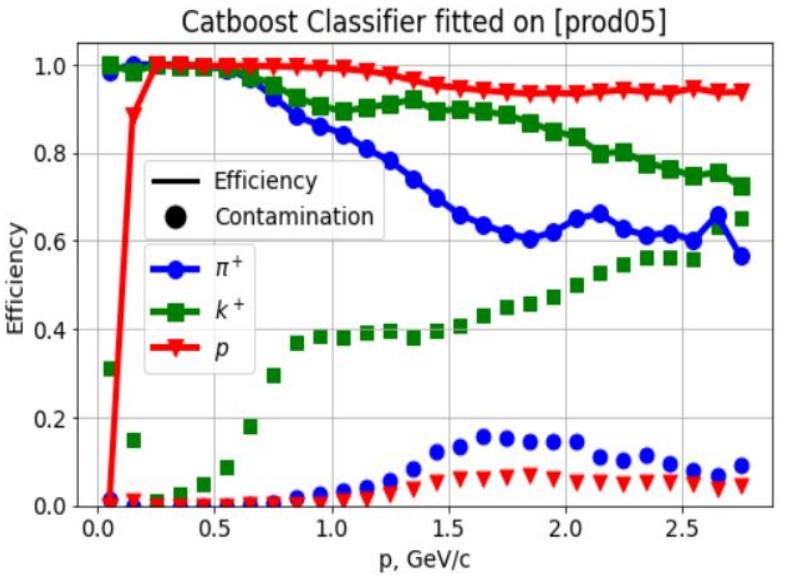
Текущие результаты

$$Efficiency = \frac{right\ identified\ tracks}{all\ tracks} \qquad Contamination = \frac{wrong\ identified\ tracks}{identified\ tracks}$$

Эффективность идентификации:

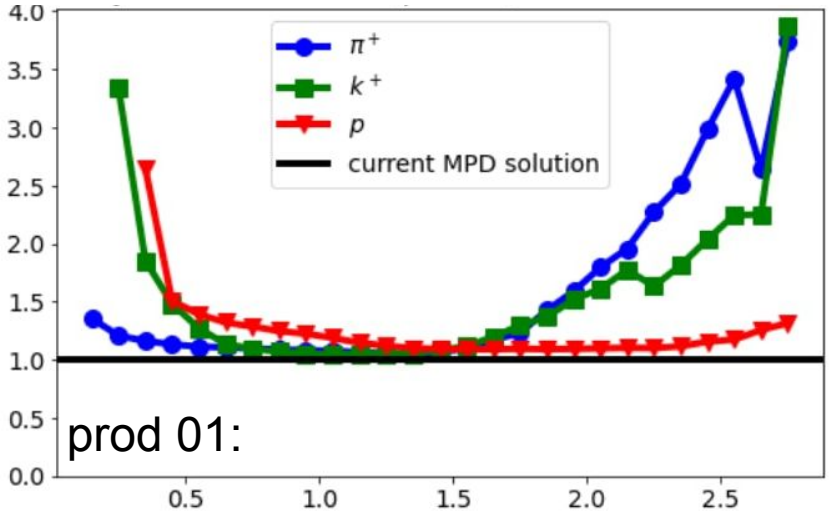


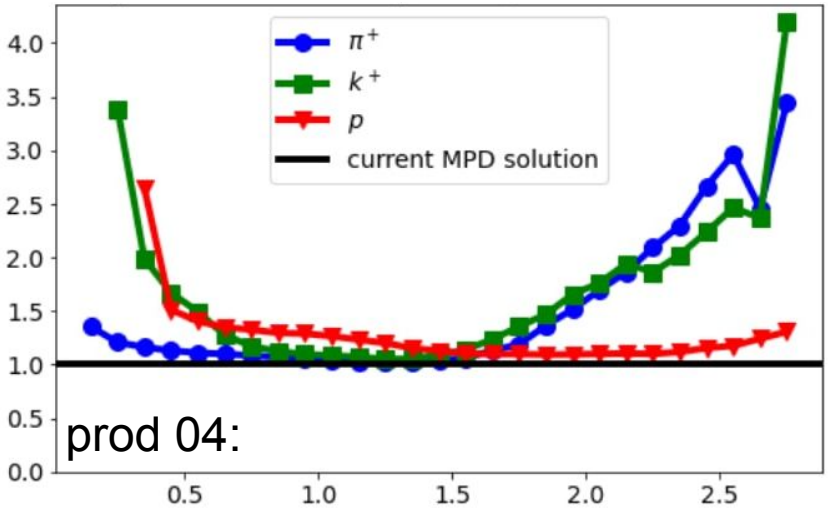


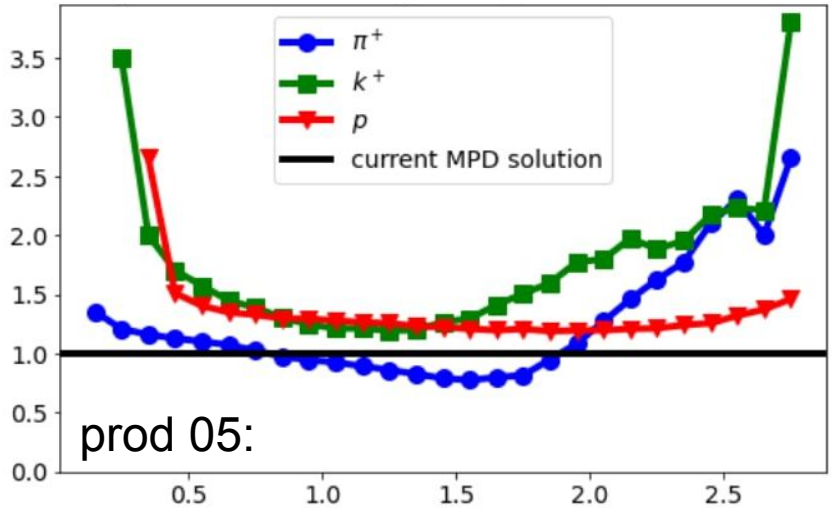


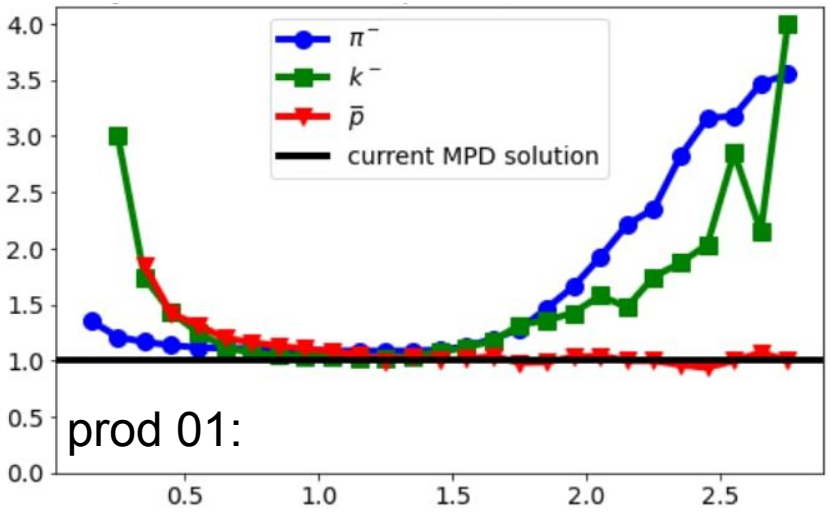
Сравнение CatBoost и текущего решения в MPD

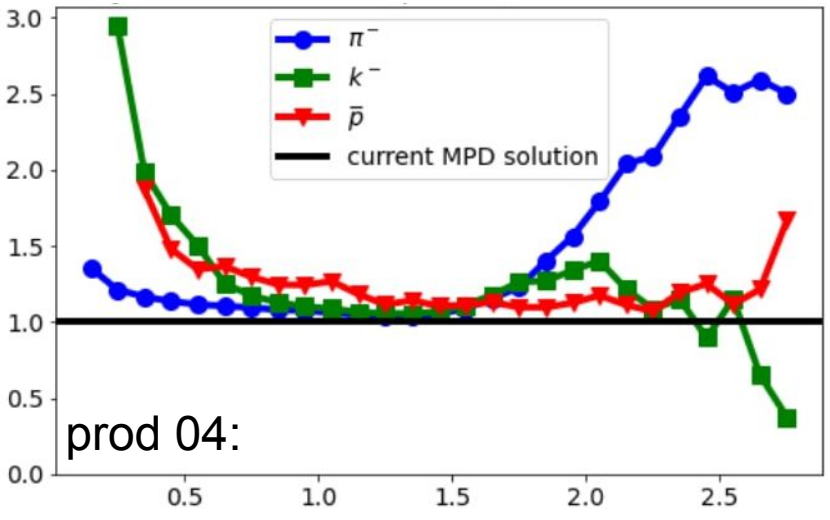
Отношения эффективностей CatBoost / N-sigma:

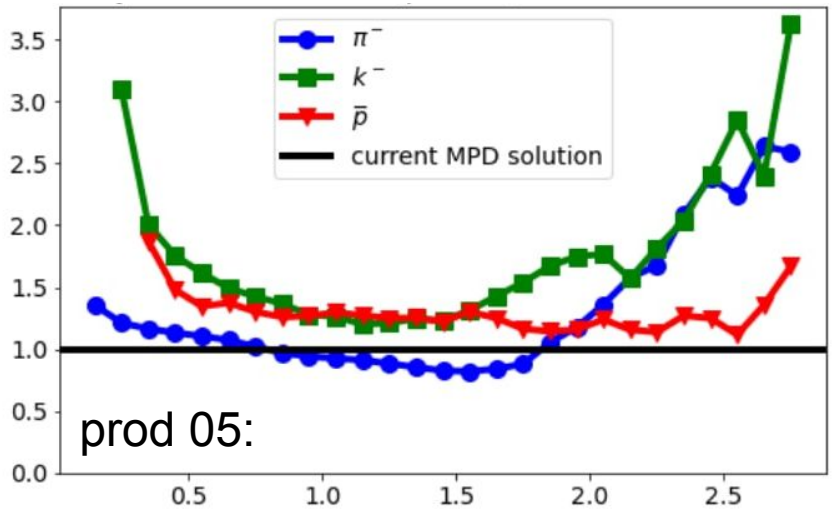












Хакатон. Команда JINR

2nd workshop on Artificial Intelligence

for the Electron Ion Collider

Oct 10 – 14, 2022

Team Name	Participant 1	Participant 2	Participant 3	Participant 4
JINR	Alexey Aparin	Artem Korobitsin	Grigorii Tolkachev	Vladimir Papoyan

The hackathon starts at 10:00AM E.T on 14th October 2022, and until 5:00PM E.T on 14th October 2022

https://indico.bnl.gov/event/16586/page/435-hackathon

Формат данных

Данные были представлены в .csv формате.

1. Общее число колонок составляло 185 штук
2. Каждой строке соответствовало одно событие с одной частицей π^+ (211) или K^+ (321)
3. Первые пять колонок включали:
    - eventID
    - PID
    - momentum [GeV]
    - theta [Θ]
    - phi [ϕ]
4. Остальные 180 колонок:
    - (X0, X1, …, X59), (Y0, Y1, …, Y59), (Z0, Z1, …, Z59) - расположение трека частицы

Формат данных

eventID	PID	momentum	theta	phi	X0	X1	X2	X3	X4	...	Z50	Z51	Z52	Z53	Z54	Z55	Z56	Z57	Z58	Z59
0	321	15.0	19.999999	0.0	1423.794456	1405.205369	1430.641544	1403.547915	1433.835355	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	321	15.0	19.999999	0.0	1408.906905	1404.078978	1414.115390	1439.424294	1403.474218	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	321	15.0	19.999999	0.0	1430.051250	1426.306782	1432.112305	1439.064895	1401.014764	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	211	15.0	19.999999	0.0	1394.185996	1402.319707	1442.940360	1392.319962	1434.855558	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	211	15.0	19.999999	0.0	1430.613166	1436.207042	1410.129402	1394.685095	1443.490183	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1499995	321	15.0	19.999999	0.0	1435.607311	1429.843666	1438.180289	1431.681709	1412.254015	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1499996	321	15.0	19.999999	0.0	1438.764633	1438.127191	1436.564538	1419.634826	1417.343095	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1499997	211	15.0	19.999999	0.0	1446.151908	1401.789510	1406.219731	1389.773612	1388.870196	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1499998	321	15.0	19.999999	0.0	1422.369126	1420.570300	1411.851823	1439.666466	1426.353037	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1499999	211	15.0	19.999999	0.0	1425.667724	1395.467156	1402.992812	1450.238262	1398.860692	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Задачи

Бинарная классификация π^+ (PID=211) или K^+ (PID=321)

1.

Training Events	1.5 Million Events	With Magnetic Field (~ 1.5T)
Momentum	15 GeV/c	at Interaction Point (0, 0, 0)
Theta θ	$20°$	at Interaction Point (0, 0, 0)
Phi ϕ	$0°$	at Interaction Point (0, 0, 0)

2.

Training Events	3 Million Events	With Magnetic Field (~ 1.5T)
Momentum	$15 - 20 GeV/c$	at Interaction Point (0, 0, 0)
Theta θ	$15 - 16°$	at Interaction Point (0, 0, 0)
Phi ϕ	$0 - 5°$	at Interaction Point (0, 0, 0)

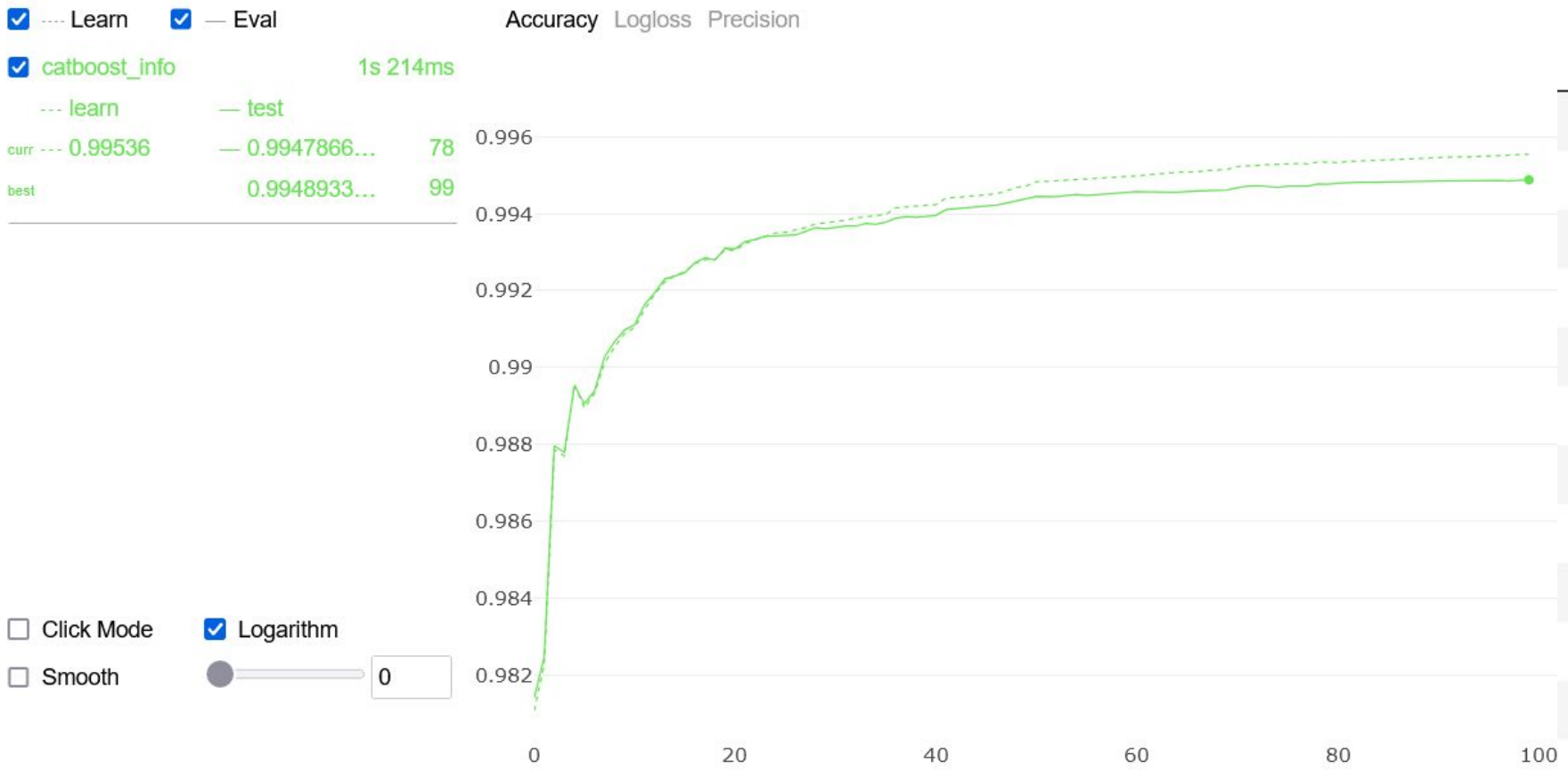
3.

Training Events	3 Million Events	With Magnetic Field (~ 1.5T)
Momentum	$15 - 20 GeV/c$	at Interaction Point (0, 0, 0)
Theta θ	$15 - 16°$	at Interaction Point (0, 0, 0)
Phi ϕ	$0 - 5°$	at Interaction Point (0, 0, 0)

в третьей задаче в каждое событие добавлены зашумленные хиты

Разведочный анализ данных

график обучения CatBoostClassifier:

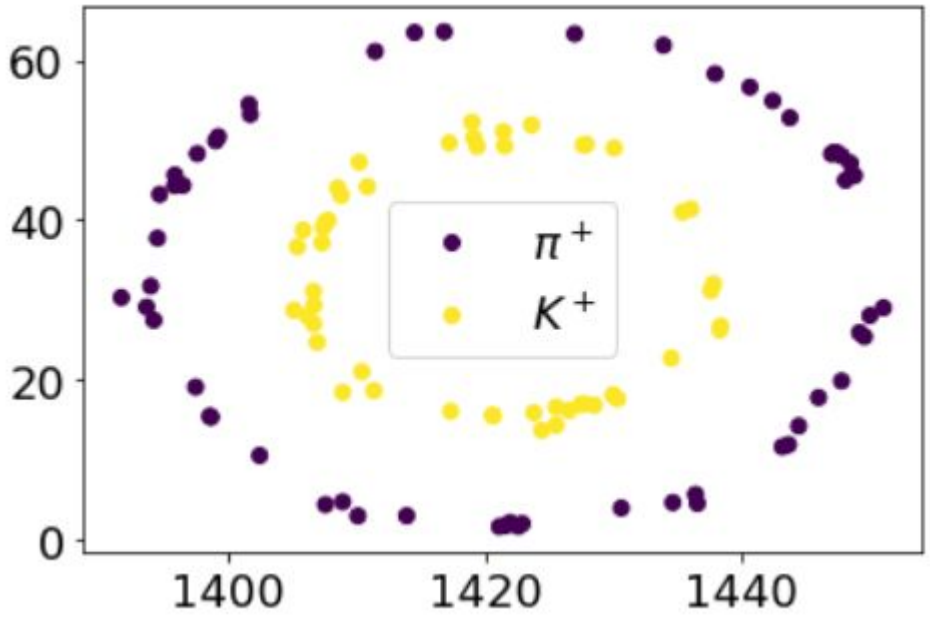


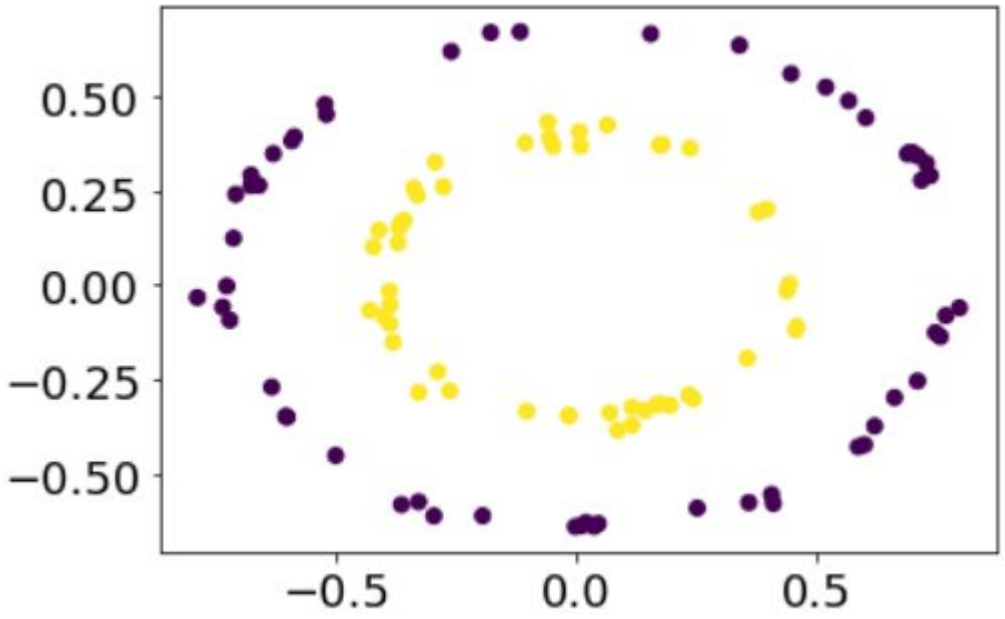
значимость координат хитов:

	Feature Id	Importances
0	Y0	15.519041
1	X0	9.689802
2	Y1	7.143211
3	Z0	5.942148
4	X1	4.285659
...	...	...
175	Z55	0.000000
176	Z56	0.000000
177	Z57	0.000000
178	Z58	0.000000
179	Z59	0.000000

Попытка сконструировать новый признак

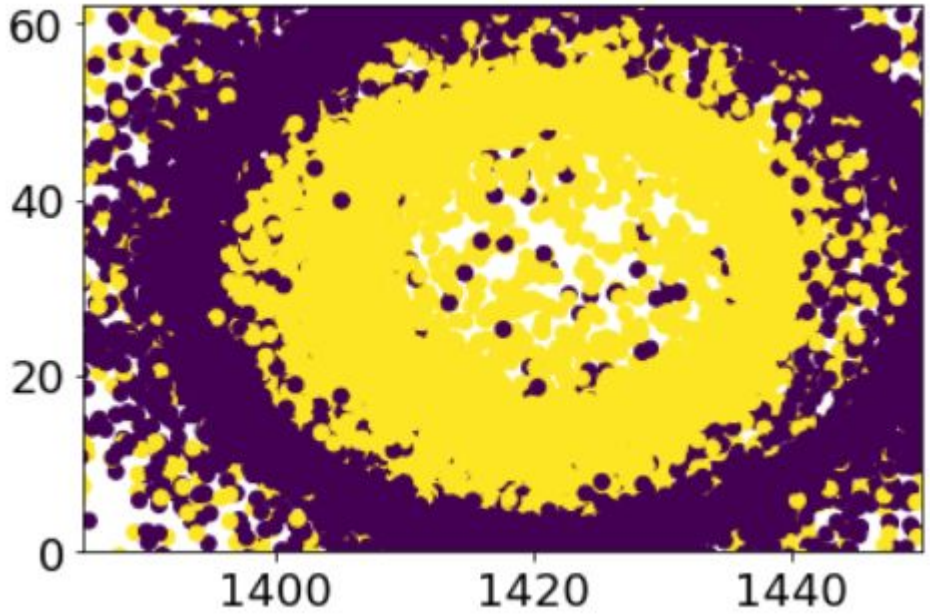
начальные точки первых 100 треков

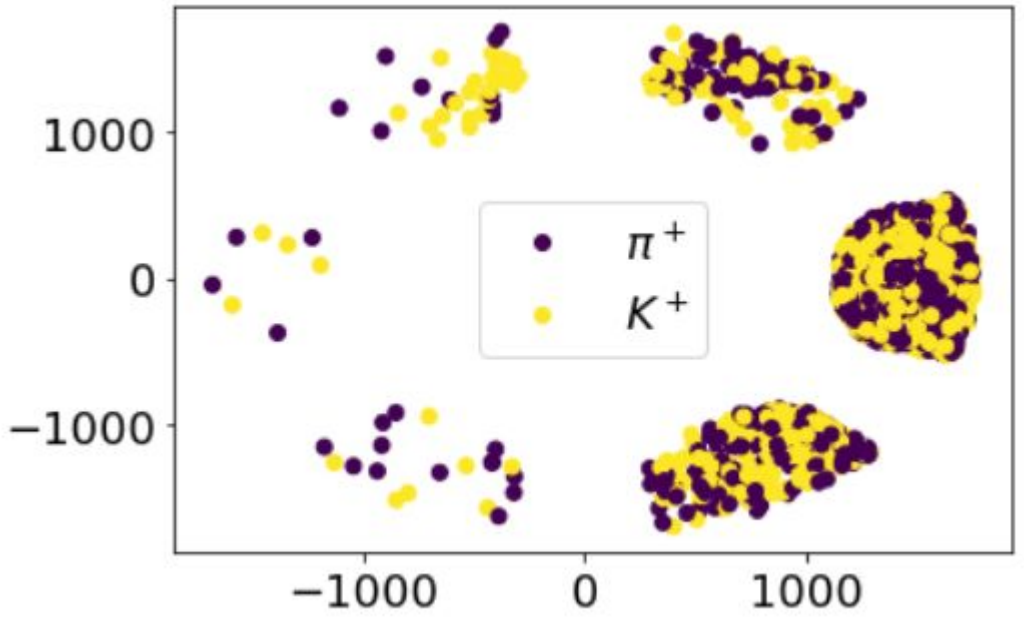




Попытка сконструировать новый признак (радиус)

начальные точки первых 500 000 треков





Точность классификации не увеличилась после добавления радиусов для каждой точки.

Начальные точки треков могут быть использованы для идентификации частиц в эксперименте MPD

Способ оценки

$$\text{ACCURACY} = \frac{\Sigma_{N_{test}} |y_{pred} == y_{true}|}{N_{test}}$$

$$\text{SCORE} = 0.0 \quad \{\text{IF ACCURACY} < \text{THRESHOLD}\}$$

$$\text{SCORE} = 50.0 + 50.0 \times \frac{(\text{ACCURACY} - \text{THRESHOLD})}{100.0 - \text{THRESHOLD}}$$

Problem Number	Threshold Accuracy
Problem 1	94%
Problem 2	86%
Problem 3	80%

AI4EIC Hackathon

Congrats Team JINR!!!!!!!! (submission on 10-14-2022)

Vladimir_P	96.692	2022-10-14 19:18:12.008951	OK
Vladimir_P	98.389	2022-10-14 19:06:44.048222	OK
Vladimir_P	98.815	2022-10-14 19:13:07.256701	OK

Backup

Важность характеристик в зависимости от обучающей выборки

Важность признаков позволяет определить насколько в среднем изменится ответ модели при изменении значения характеристики частицы [1]

prod01:

	Feature Id	Importances
0	charge	48.976478
1	p	15.612522
2	m2	13.219858
3	dedx	12.504383
4	dca	2.931781
5	nHits	2.682914
6	eta	1.732293
7	Vz	0.904500
8	Vx	0.757425
9	Vy	0.677845

prod04:

	Feature Id	Importances
0	charge	52.595520
1	p	16.143578
2	m2	11.179546
3	dedx	9.959441
4	eta	3.202594
5	dca	3.178775
6	nHits	2.890517
7	Vy	0.322261
8	Vx	0.293670
9	Vz	0.234098

prod05:

	Feature Id	Importances
0	charge	43.753433
1	p	19.143319
2	dedx	18.371532
3	m2	9.106441
4	dca	3.549774
5	nHits	2.178229
6	eta	1.912249
7	Vz	0.802412
8	Vx	0.630954
9	Vy	0.551657

Чем выше значение **важности характеристики** частицы, тем больше в среднем изменится значение ответа модели, в случае изменения самой характеристики

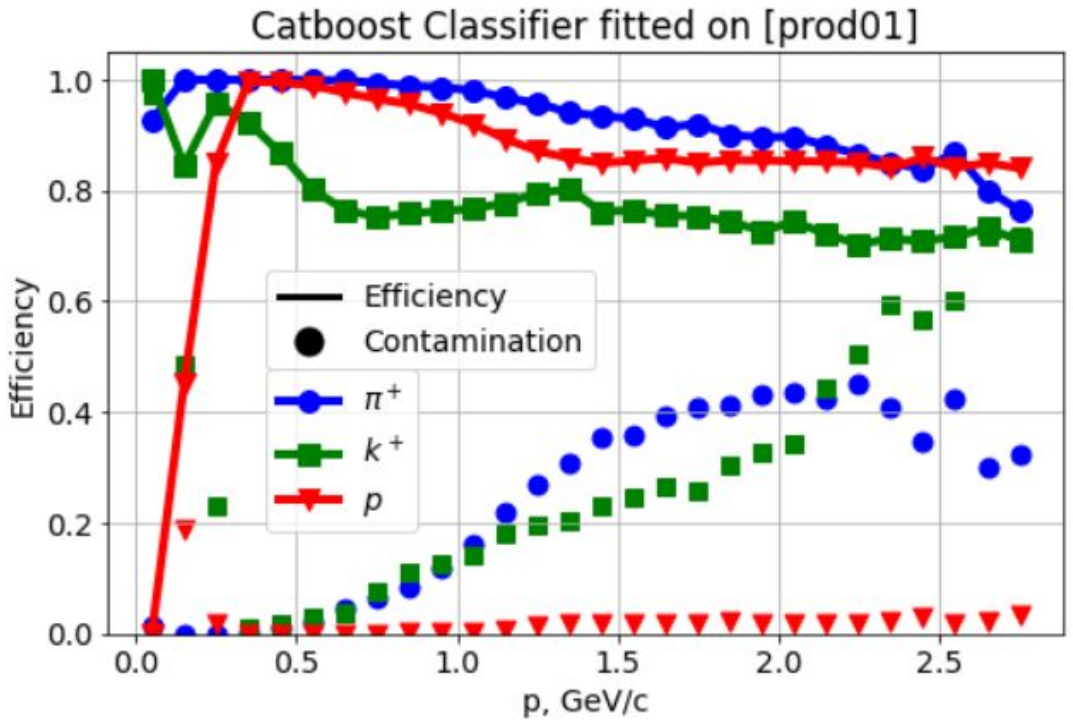
[1] https://catboost.ai/en/docs/concepts/fstr#regular-feature-importance

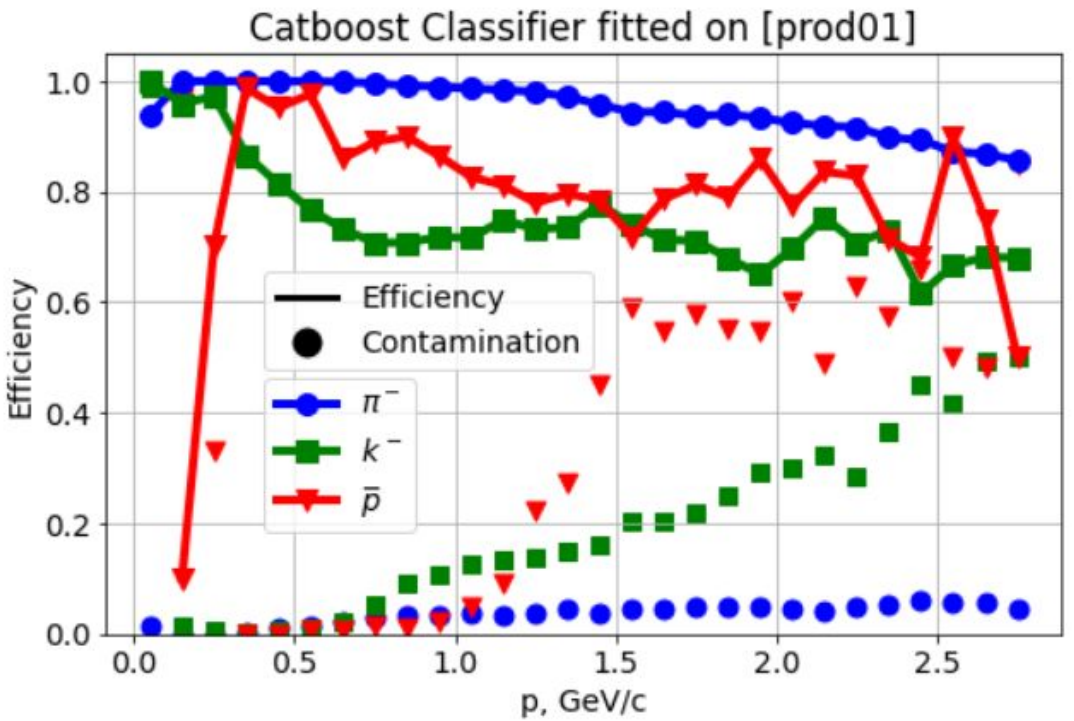
Текущие результаты

$$Efficiency = \frac{right\ identified\ tracks}{all\ tracks}$$

$$Contamination = \frac{wrong\ identified\ tracks}{identified\ tracks}$$

Эффективность идентификации:



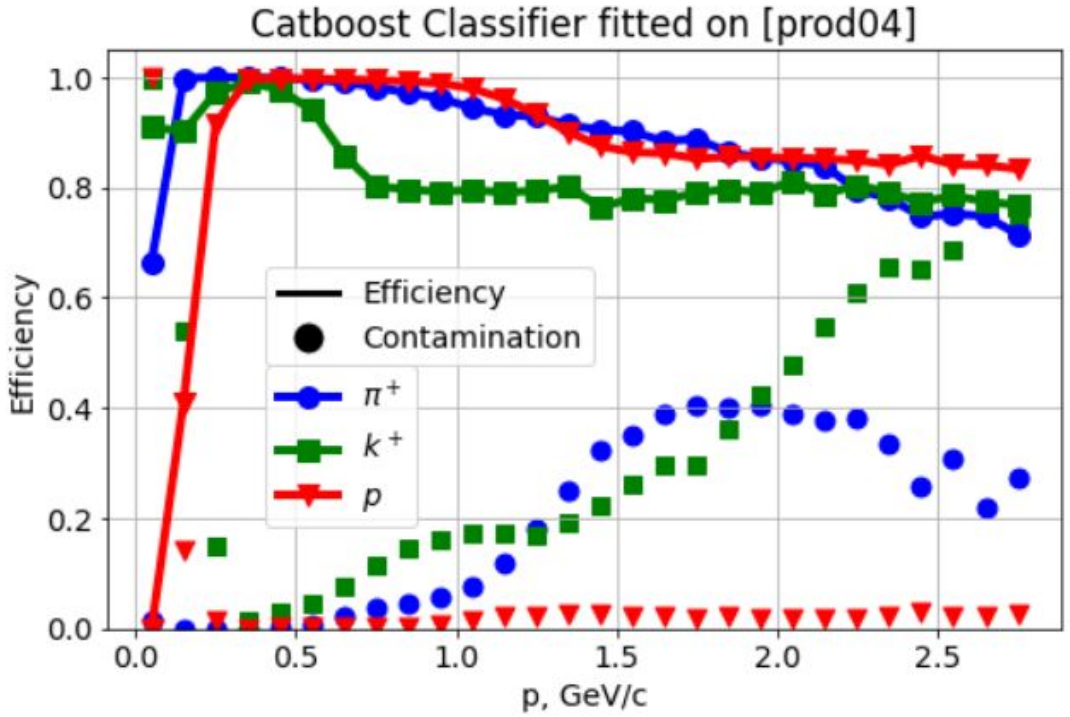


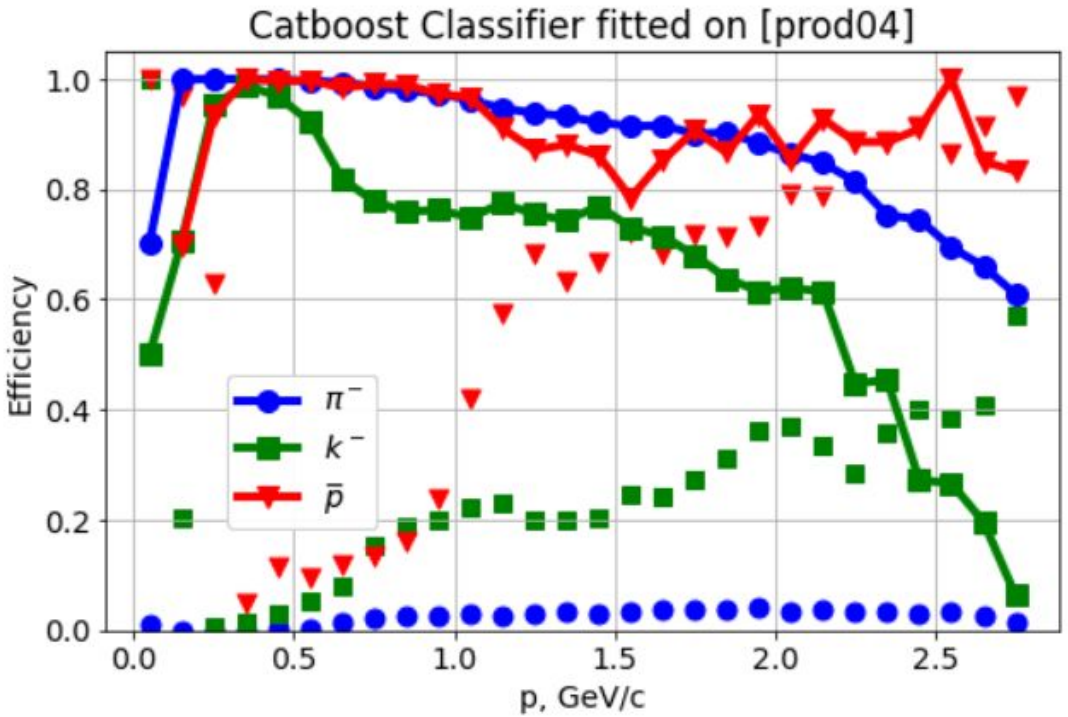
Текущие результаты

$$Efficiency = \frac{right\ identified\ tracks}{all\ tracks}$$

$$Contamination = \frac{wrong\ identified\ tracks}{identified\ tracks}$$

Эффективность идентификации:



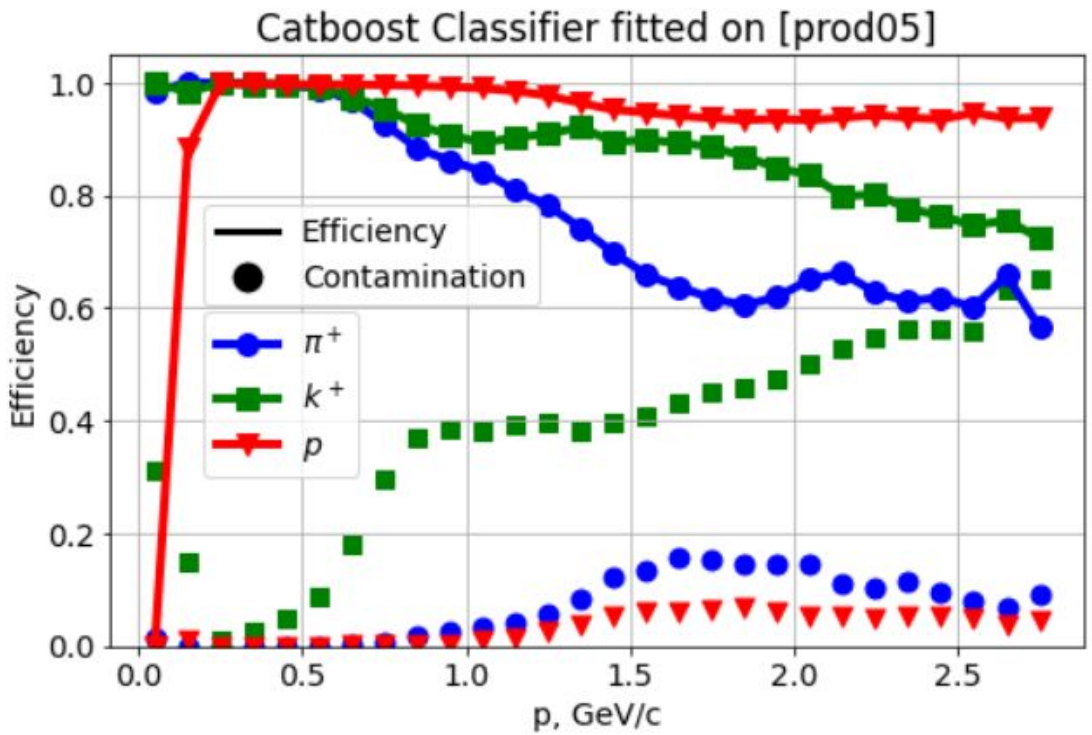


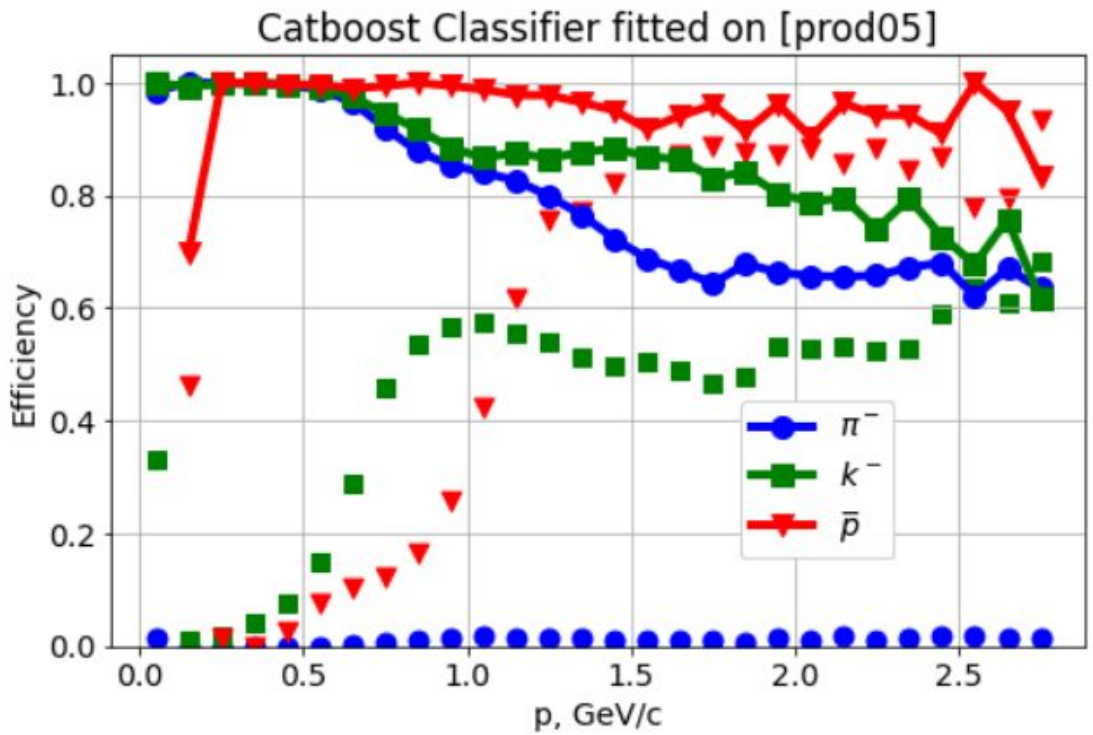
Текущие результаты

$$Efficiency = \frac{right\ identified\ tracks}{all\ tracks}$$

$$Contamination = \frac{wrong\ identified\ tracks}{identified\ tracks}$$

Эффективность идентификации:





Текущие результаты

$$Efficiency = \frac{right\ identified\ tracks}{all\ tracks} \qquad Contamination = \frac{wrong\ identified\ tracks}{identified\ tracks}$$

Эффективность идентификации: